OXFORD

Briefings in Bioinformatics, 2022, **23(6)**, 1–14

https://doi.org/10.1093/bib/bbac380

Advance access publication date 12 September 2022

Problem Solving Protocol

MCIBox: a toolkit for single-molecule multi-way chromatin interaction visualization and micro-domains identification

Simon Zhongyuan Tian, Guoliang Li, Duo Ning, Kai Jing, Yewen Xu, Yang Yang, Melissa J. Fullwood, Pengfei Yin, Guangyu Huang, Dariusz Plewczynski, Jixian Zhai, Ziwei Dai, Wei Chen and Meizhen Zheng

Corresponding authors. Meizhen Zheng, Shenzhen Key Laboratory of Gene Regulation and Systems Biology, Department of Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen 518055, China. Tel.: +86 755 88018669; Fax.: +86 755 88018669; E-mail: zhengmz@sustech.edu.cn; Wei Chen, Shenzhen Key Laboratory of Gene Regulation and Systems Biology, Department of Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen 518055, China. Tel.: +86 88018449; Fax.: +86 88018449; E-mail: chenw@sustech.edu.cn; Simon Zhongyuan Tian, Shenzhen Key Laboratory of Gene Regulation and Systems Biology, Department of Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen 518055, China. Tel.: +86 755 88018669; Fax.: +86 755 88018669; E-mail: tianzy3@sustech.edu.cn

Abstract

The emerging ligation-free three-dimensional (3D) genome mapping technologies can identify multiplex chromatin interactions with single-molecule precision. These technologies not only offer new insight into high-dimensional chromatin organization and gene regulation, but also introduce new challenges in data visualization and analysis. To overcome these challenges, we developed MCIBox, a toolkit for multi-way chromatin interaction (MCI) analysis, including a visualization tool and a platform for identifying micro-domains with clustered single-molecule chromatin complexes. MCIBox is based on various clustering algorithms integrated with dimensionality reduction methods that can display multiplex chromatin interactions at single-molecule level, allowing users to explore chromatin extrusion patterns and super-enhancers regulation modes in transcription, and to identify single-molecule chromatin complexes that are clustered into micro-domains. Furthermore, MCIBox incorporates a two-dimensional kernel density estimation algorithm to identify

Simon Zhongyuan Tian received his PhD degree in Systems Biology from Yamaguchi University, Japan in 2014. He is a research assistant professor in the School of Life Sciences, Southern University of Science and Technology, China. His research interests include developing 3D genome mapping technology and new tools for data processing, analysis and visualization.

Guoliang Li received his PhD degree in computer science from the National University of Singapore, Singapore in 2009. He is a professor in bioinformatics in the College of Informatics, Huazhong Agricultural University, China. His research interests include the development and applications of 3D genomics methods.

Duo Ning PhD candidate in the School of Life Sciences, Southern University of Science and Technology, China. He focuses on applying long-reads ChIA-PET method on investigating gene regulation in cancer.

Kai Jing is a master student in the School of Life Sciences, Southern University of Science and Technology, China. He focuses on multi-omics data processing and analyzing.

Yewen Xu is a master student in the School of Life Sciences, Southern University of Science and Technology, China. She focuses on the development of 3D genomics methods.

Yang Yang PhD candidate in the School of Life Sciences, Southern University of Science and Technology, China. She focuses on the development of 3D genomics methods.

Melissa J. Fullwood received the BSc degree (Hons.) from Stanford University, USA, and the PhD degree from the National University of Singapore, Singapore, in 2005 and 2009, respectively. She is a principal investigator in the Cancer Science Institute of Singapore, National University of Singapore, an assistant professor in the School of Biological Sciences, Nanyang Technological University and an adjunct principal investigator in the Institute of Molecular and Cell Biology, A*STAR, Singapore. Her research interests include investigating 3D genome organization in cancer.

Pengfei Yin is a master student in the School of Life Sciences, Southern University of Science and Technology, China. He focuses on applying deep learning to investigate chromatin structures.

Guangyu Huang is a master student in the School of Life Sciences, Southern University of Science and Technology, China. He focuses on chromatin interaction in single cell.

Dariusz Plewczynski is the head of Laboratory of Bioinformatics and Computational Genomics at the Faculty of Mathematics and Information Science, Warsaw University of Technology, and the Laboratory of Functional and Structural Genomics at the Centre of New Technologies, University of Warsaw, Poland. His interests are focused on functional and structural genomics, exploring the population variability of human DNA sequence, chromatin structure and its biological function at the whole genome scale. His statistical learning, data driven attempts make use of the vast wealth of experimental data produced by high-throughput genomics projects.

Jixian Zhai associate professor in the School of Life Sciences, Southern University of Science and Technology, China. His research interests include high-throughput sequencing and epigenomics.

Ziwei Dai principal investigator and assistant professor in the School of Life Sciences, Southern University of Science and Technology, China. Her research focuses on quantitative biology of cancer metabolism.

Wei Chen is a professor of Systems Biology and Chair of Department of Biology, School of Life Sciences, Southern University of Science and Technology, China. The research interest of his lab is to understand the mechanisms underlying posttranscriptional gene regulation and their important role in causing human diseases as well as shaping regulatory divergence between different organisms during evolution.

Meizhen Zheng is a principal investigator and assistant professor in the School of Life Sciences, Southern University of Science and Technology, China. Her research interests include the development of 3D genome mapping technology with single-molecule precision in single nuclei and development of spatial genomics methods in single nuclei.

Received: May 15, 2022. **Revised:** August 5, 2022. **Accepted:** August 9, 2022

micro-domains boundaries automatically. These micro-domains were stratified with distinctive signatures of transcription activity and contained different cell-cycle-associated genes. Taken together, MCIBox represents an invaluable tool for the study of multiple chromatin interactions and inaugurates a previously unappreciated view of 3D genome structure.

Keywords: multi-way chromatin interactions, single-molecule, micro-domains, visualization, machine learning, chromatin extrusion, super-enhancers

Introduction

Eukaryotic genomes are organized into a three-dimensional (3D) multiscale structure within the nucleus, and this structure is often associated with gene regulation [1]. The proximity ligation-based 3D genome mapping technologies, such as Hi-C and ChIA-PET, have enabled global mapping of chromatin interactions and characterization of nuclear genome organization in multiple scales, ranging from loops, topologically associated domains (TADs), to A/B compartments [2–6]. State-of-the-art visual tools such as Juicebox, HiGlass, HiCPlotter, 3D Genome Browser, WashU Epigenome Browser, Delta, BASIC Browser and Nucleome Browser have been developed for users to explore 3D genome organization and long-range chromatin interactions conveniently [7–14]. Thereinto, BASIC Browser displays RNA polymerase II (RNAPII)-mediated promoter–enhancer contacts as complex chromatin interactions [3], i.e. RAIDs (RNAPII-mediated chromatin interaction domains), which offers the genomics evidence for the proposed model of 'transcription factories' from RNAPII foci observed via imaging data [15]. These tools have provided deep insight into the principle of gene regulation.

However, the aforementioned experimental methods and visualization tools can only capture pairwise interactions. Moreover, these methods present accumulated results from bulk cell populations; thus, they cannot precisely locate the interacting loci with high heterogeneity in single nuclei. Therefore, it remains unclear whether these multiplex chromatin interactions happen in individual nuclei or composed by different single-molecule chromatin complexes from different nuclei and finally showing as accumulated results. Thus, these methods cannot reveal the detailed nature of chromatin contacts for answering questions regarding the modes for TADs in individual nuclei or the composed pattern of super-enhancers associated with key genes for cell destination.

In order to overcome these limitations, a few ligation-free technologies have been developed. GAM, by sequencing a collection of thin cryosectioned nuclear profiles for genome architecture mapping [16], revealed three-way contacts among super-enhancers for specific gene regulation. SPRITE, a split-pool recognition of chromatin interactions by tag extension [17], showed 'active hub' around nuclear speckles enriched with RNAPII activity and 'inactive hub' around nucleolus enriched rDNA corresponding to low transcribed activity. ChIA-Drop, which is the microfluidic-based and barcode-linked sequencing for chromatin interaction analysis [18], showed a strong directionality bias towards the gene body for supporting one-sided extrusion model of transcription. These methods have enabled users to capture multiple interacting genomic loci simultaneously.

ChIA-Drop, SPRITE and GAM approaches have demonstrated the multi-way chromatin interactions (MCI) involved in the 3D genome architecture and gene regulation and comprehensively mapped chromatin contacts at a previously unappreciated level [19]. ChIA-Drop data further provided evidence that the single-molecule chromatin complexes with high heterogeneity and also clustered together as microscale domains via certain similarity, contributing to the chromatin structure such as TADs and RAIDs as accumulated results [18]. The emerging evidence indicates that chromatin microscale domains that are organized into TADs exhibit higher tendency of cell-type-specific 3D genome structure than low-resolution TADs [20]. These ligation-free methods including SPRITE and ChIA-Drop can elucidate principles of genome folding at microscale by clustering their similarity and can demultiplex the averaged interacting loci precisely into single nuclei multi-way interactions with single-molecule precision.

When investigating the ligation-free data of ChIA-Drop, GAM and SPRITE, we found that individual chromatin TADs contained several micro-domains. In this study, for the microscale domain constructed into TADs were called microTADs, and that constructed into RAIDs associated with transcription factories were called microTFY, respectively, for further investigation.

So far, these novel exciting findings were identified with computational methods and integrative analysis. However, the visualization tools for real-time detailed profiling of single-molecule chromatin complexes and the characterization of micro-domains are currently lacking. Here we report MCIBox, a new toolkit for profiling multi-way chromatin interactions and visualizing micro-domains clustered from single-molecule chromatin interaction complexes. These approaches include the hierarchical clustering algorithm on data matrix from GAM, SPRITE and ChIA-Drop datasets, respectively, or various clustering algorithms on scatter plot matrix created with dimensionality reduction on that data matrix.

In addition, we added a program in MCIBox for micro-domain characterization, a two-dimensional kernel density estimation (2D KDE) contour map-based 'MCI-2kde'. By focusing on an RNAPII-associated ChIA-Drop dataset of *Drosophila* S2 cells [18], we applied MCI-2kde program to automatically determine the boundary of micro-domains, which are micro-transcription factories (microTFY) mediated by RNAPII transcription factors (TF). We identified 578 microTFY from the 126 of 476 RAIDs at length scales above 150 kb that determined by previously pairwise RNAPII ChIA-PET data [18]. These microTFY were stratified from various patterns with distinctive signatures of transcription activity and histone modification. Interestingly, different microTFY in a RAID contained specific genes in the different phases of a cell cycle, indicating that the application of MCIBox allows us to distinguish single-molecule chromatin interactions with cell-cycle specificity even among a same cell line.

Taken together, MCIBox not only can real-time explore the chromatin topological structure and higher-order chromatin organization in multi-way chromatin interactions data, such as the chromatin extrusion patterns in gene transcription and chromatin organization, but also can characterize the similarity of these single-molecule chromatin complexes by clustering them into microscale domains.

Methods

MCIBox includes MCI-view and MCI-2kde

MCIBox toolset realizes two main functions, the function for single-molecule or single-cell data clustering visualization is realized in MCI-view module and the function to define the

boundaries of micro-domains such as microTFY is done by MCI-2kde module. MCIBox is programmed in the framework of R Shiny Server (https://shiny.rstudio.com), which is convenience to build an interactive web application directly from R script. Without special description, R packages used in this work are from CRAN package repository (https://cran.r-project.org), where we can find source codes and description documents.

Data preparation for MCI-view

Currently, MCI-view main data are sourced from 'ligation-free' new generation 3D genome techniques: ChIA-Drop, SPRITE and GAM (Figure 1A and B), which capable to capture multiplex interactive fragments from a shared chromatin complex (Figure 1C). Because the data format of these techniques out of their pipelines are different, we programmed interface tools to transform these data into a uniform RGN format that MCI-view required. A RGN file actually is a fragment list, in which each data line containing a fragment with four basic columns (chromosome, start, end and barcode) representing fragment genomic region coordinates and with a complex identity (complex barcode ID). Interface tools transform each source data into a set of RGN files composed by one chromosome per file for easier accessibility. All RGN files used for MCI-view are stored in RDS format, a binary format of R objects with a quick data loading speed and smaller storage space. For user convenience, we have preprocessed several datasets for MCI-view, such as: ChIA-Drop (dm3), RNAPII ChIA-Drop (dm3), SPRITE (hg19) and GAM (mm9), etc. Users can visualize them by simply click the 'FIN' button on the top-left of MCI-view web-app to select the interested data from the list, and subsequently the selected data will be loaded automatically.

MCI-view workflow and visualization tracks

MCI-view is the visualization module of MCIBox toolkit and briefly includes three layers: calculation layer, basic-view layer and extensive-view layer (Supplementary Figure S1). The calculation layer executes the main algorithms computation functions, besides clustering algorithms and dimensionality reduction algorithms. The basic-view layer contains the visualization plots that arise directly out of the calculation layer, such as Fragment-view. The extensive-view layer performs some filtering operations upon the illustrated data, for example 'Target loci' for DNA-FISH Probe-based loci filtering.

MCI-view for data visualization includes four main steps: (i) **Data Input Step**: At the very beginning, MCI-view adopts data of an interested genomic region from the according chromosome RGN file of the source library selected. (ii) **Data Format Step**: Then the input data is integrated as barcoded linked-reads and binned to matrix, formed as rows are barcodes representing complexes, columns are (original or binned) genomic region (Figure 1D and E). There are also three types of displays that based on data accumulation of bin–bin or bins, could be generated in this step: 2D heatmap, 2D loop or 1D coverage (Supplementary Figure S1A–C). (iii) **Data Clustering Step**: MCI-view performs clustering function in either of the two strategies. The first strategy is clustering high-dimensional (HD) data, which directly runs hierarchical clustering upon the rows of the HD matrix constructed from the former step (Figure 1F). HD Cluster-view and HD Fragment-view are for viewing clustered HD data (Supplementary Figure S1D and E). The second strategy is clustering low-dimensional data (LD, Figure 1G) that runs clustering after dimensionality reduction by assembling seven dimensionality reduction algorithms and seven clustering algorithms comprehensively. In the second strategy, MCI-view supplies selection of the combination of two types of algorithms e.g. Uniform Manifold Approximation and Projection (UMAP) plus Hk-means, following the principle of obtaining proper separated groups with unique colors denoted by clusters in the LD-clustering scatter plot (Supplementary Figure S1F). LD-clustering scatter plot, LD Cluster-view and LD Fragment-view are tracks for clustering LD data visualization (Supplementary Figure S1F–H). (iv) **Data Filtering Step**: MCI-view supplies functions for users to select data by genomic locations, such as: 'Target loci' for DNA-FISH Probe-based loci filtering, 'Chromatin organization pattern' for CTCF Motif-based loci filtering, 'Transcription pattern' for Promoter-based loci filtering and 'Transcription regulation pattern' for Super-enhancers loci filtering etc. (Supplementary Figure S1I–K). Users can also filter data by selecting out un-interested clusters to achieve a clean 'Mask-based' data view (Supplementary Figure S1L).

We prepared a user manual with demo videos in the MCIBox GitHub repository (https://github.com/ZhengmzLab/MCIBox) for guiding the installation and utilization of MCI-view.

MCI-view clustering strategy

The MCI-view clustering strategy includes two ways: the 1st way is the HD-clustering module, the 2nd way is the LD-clustering module. HD-clustering module employs Hierarchical clustering algorithm directly upon genomic data matrix of a higher dimension, by default, a given genomic region is divided into 40 bins allow for a faster-speed in clustering. LD-clustering module employs a two-step strategy to reorganize complexes arrangement, it performs a dimensionality reduction algorithm on a data matrix with 40 bins (i.e. with 40 features or dimensions), projects it into 2D data (i.e. LD data) and immediately executes a clustering algorithm. In LD-clustering module, we totally integrated seven dimensionality reduction algorithms and seven clustering algorithms, and users can freely select suitable algorithms combination for a specific data. When using MCI-view, commonly LD-clustering procedures can get better clustering effect than that of HD-clustering, with a compromise of being more time-consuming (3- to 5-folds). We integrate silhouette score in MCI-view for evaluating the quality of clustering, which allows users to select a suitable clustering by the silhouette score ratings [21].

MCI-view implemented clustering algorithms

Clustering algorithms are unsupervised machine learning algorithms that aim to classify a set of unlabeled data into several groups, referred to as 'clusters' [22]. Data points show more similarity to each other within a same cluster and more dissimilarity between different clusters. One aim of the development of MCI-view is to code a software to visualize different groupings of multi-way contacts data, such as ChIA-Drop, SPRITE, GAM and other single-cell omics data. Here, we integrated different kinds of clustering algorithms capable of cluster different types of data. These clustering algorithms were described below in briefly (The details and default parameters used in MCIBox were described in Supplementary Data).

- Hierarchical clustering [23] belongs to connectivity-based models, which is based on distance connectivity.
- K-means clustering [24] belongs to centroid-based models, which defines each cluster with a single mean vector.
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [25] belongs to density-based models, which models clusters from a connected dense region.
- Gaussian Mixture Models (GMM) clustering algorithm [26] belongs to distribution-based models, which defines clusters

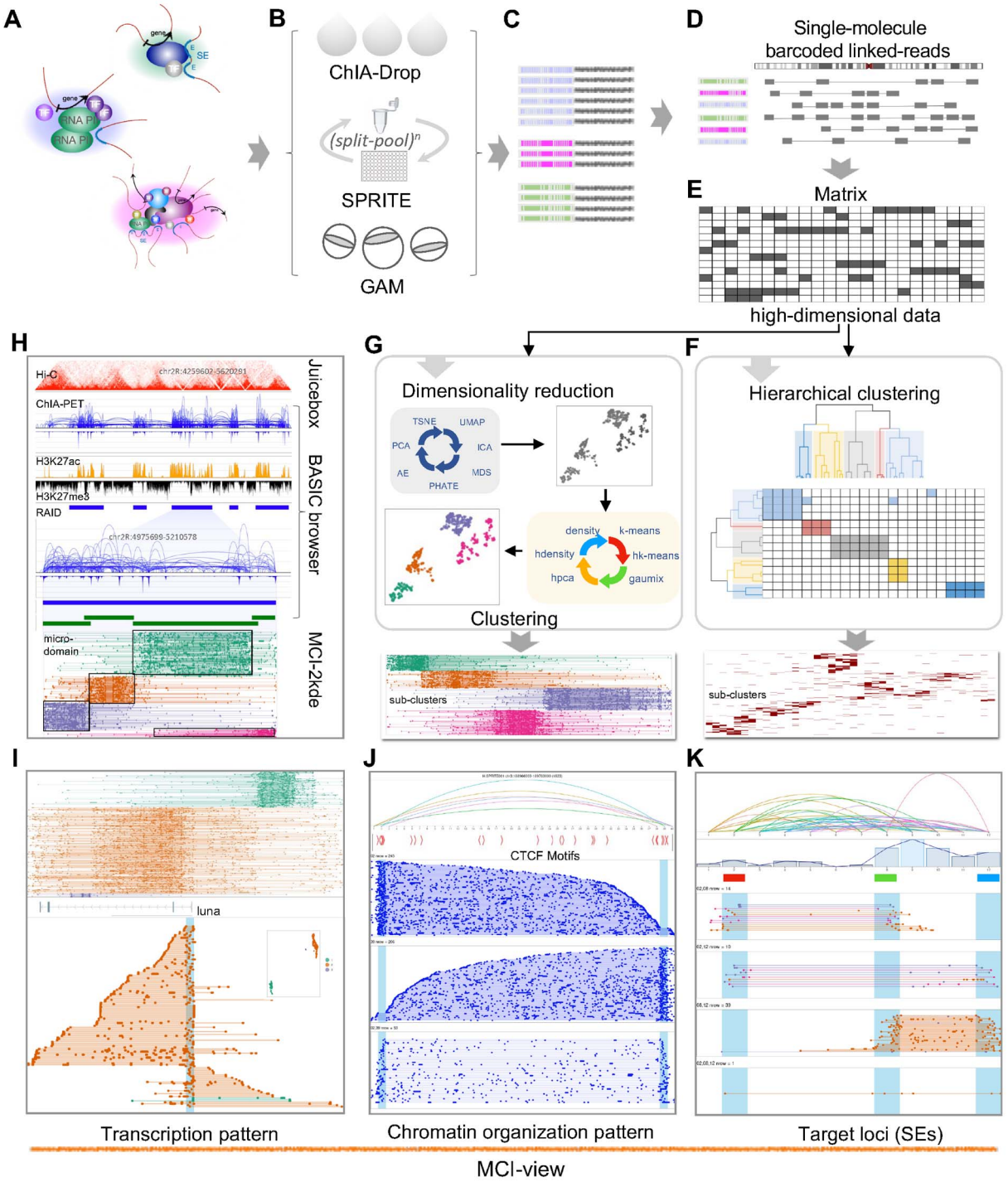

Figure 1. Overview of MCIBox for multi-way chromatin interactions analysis. (**A**–**C**) Schematic for generating multi-way chromatin interactions datasets with different ligation-free based methods including ChIA-Drop, SPRITE and GAM. (**A**) for chromatin complexes preparation, (**B**) for different methods and (**C**) for sequencing reads that were grouped by barcodes. (**D**) A multiplex chromatin interaction complex is defined as the fragments with the same barcodes. (**E**) The chromatin complexes from the genomic region of interest were used as the input into MCIBox for matrix construction. Binned genomic regions are as columns along x-axis and binned fragments from the complexes are arranged as rows along y-axis. (**F**) Subsequently, these binned fragments were arranged by similarity via hierarchical clustering upon HD data, which are ready for Cluster-view (a visualization function in MCIBox). (**G**) Using dimensionality reduction algorithms (UMAP by default), the matrix was converted to 2D scatter plot, among them the tethered closely dots were clustered together by default algorithms of hierarchical k-means (Hk-means) for Fragment-view (a visualization function in MCIBox). (**H**) A screenshot showing micro-domain from multi-way contacts with MCI-view, along with 2D heatmap of Hi-C with Juicebox, and pairwise interaction loops and coverage from ChIA-PET with BASIC Browser. The blue bar indicates RAID, and the green bar indicates micro-domain (microTAD or microTFY). (**I**–**K**) Screenshots showing 'Transcription pattern', 'Chromatin organization pattern' and 'Target loci' for multi-way contacts with MCI-view, respectively.

using statistical distributions. GMM is a probabilistic model that used for denoting normally distributed subpopulations within an overall population.

- Fuzzy clustering [27], also known as soft clustering, is a form of clustering in which each data point can belong to more than one cluster, which was implemented to study the data points that lie equally distant from each other.
- The other three types of hybrid clustering algorithms that MCI-view used, may take advantages from both combined parts. Such as Hierarchical Clustering on Principal Components (HCPC) [28], Hierarchical k-means clustering (Hk-means) [29] and Hierarchical DBSCAN (HDBSCAN) [30]. They use the complement between the hierarchical clustering and another clustering algorithm to better highlight the main features of the data set.

MCI-view implemented algorithms of dimensionality reduction

MCI-view can perform clustering calculations directly upon data matrix with high dimensions (HD-clustering) or after a dimensionality reduction pre-step (LD-clustering). Via HD-clustering, users can get clustering results rapidly, but often compromised with overfitted and poor performance, commonly known as the curse of dimensionality refers to the situation when the data has too many features. Dimensionality reduction techniques convert the higher dimensions dataset into lesser dimensions dataset with minimal loss of information. Dimensionality reduction is widely used in the fields for dealing with high-dimensional data, such as data visualization, speech recognition, signal processing, noise reduction, cluster analysis [31].

In total, MCI-view assembled seven dimensionality reduction algorithms before carrying out clustering algorithms for better groupings and provided an interactive interface allowing users to select suitable algorithms. Dimensionality reduction algorithms mainly includes linear methods and nonlinear methods [32]. They use different strategies to project the original data onto an low-dimensional space, linearly or nonlinearly. The linear methods including linear Principal Component Analysis (PCA) [33] and Independent Component Analysis (ICA) [34] are preferred for gaussian distributed samples and non-gaussian data respectively. The nonlinear dimensionality reduction algorithms in MCI-view are described below:

- Multi-Dimension Scaling algorithm (MDS) [35] is a distance-preserving manifold learning method.
- *t*-Distributed Stochastic Neighbor Embedding (TSNE) algorithm [36] and UMAP algorithm [37] are two prevailing dimensionality reduction algorithms. TSNE considers the similarity between the local gaussian distributions of high-dimensional data and *t*-distribution of low-dimensional space, whereas UMAP constructs a graph representing the high-dimensional data then optimizes to an low-dimensional structurally similar graph. TSNE is often better at preserving the local structure, whereas UMAP is more to the global structure.
- AutoEncoder algorithm (AE) [38] is a traditional dimensionality reduction algorithm, which is an artificial neural model that involves encoder and decoder model of unsupervised learning.
- Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE) [39] is a state-of-the-art dimensionality reduction algorithm specially designed for single-cell RNA-seq, which calculates local affinities between cells first, and then the affinities are used to define transitional probabilities and spread them by a Markovian diffusion framework over the data.

Evaluation of the quality of clusters created using clustering algorithms and dimensionality reduction algorithms

Silhouette analysis is used to evaluate the appropriateness of the partitioning of a datapoint to one cluster rather than to other clusters. Silhouette score [21] is used to measure both intracluster compactness and intercluster separation. A silhouette score is in a range [-1, 1], of which a higher value represents a better clustering result.

MCI-view not only integrates a number of algorithms combinations of dimensionality reduction and clustering in its LD-clustering module, but also introduces a function of calculating the mean of silhouette score for all datapoints within a genomic region, which enable users to select algorithms with well-separated clusters by higher silhouette score. Generally, a higher silhouette score (>0.50) can be seen as a signal of having achieved a reasonable clustering result [22]. Users can also adjust parameters of algorithms in MCI-view according to the characteristics and types of their specific data to obtain better grouping results.

In this study, MCI-view uses UMAP-Hk-means combinatory algorithms as the default method for LD-clustering, since they got the highest silhouette score among the combinations of dimensionality reduction algorithms and clustering algorithms upon datasets within the RAIDs genomic regions for achieving well groupings of microTFY.

MCI-2kde defines microTFY boundary

The histogram is one of the most popular types of plots to show data distribution, which is based on probability density functions (PDF). Univariate KDE approach can generate an empirical estimate of the true PDF, without distribution forms assumption and any parameters [40]. Bivariate KDE, also known as 2D KDE [40], determines information about the joint occurrence of two related variables, which denote genomic bins (x-axis) and complexes (y-axis) as the two dimensions in this work.

To automatically obtain the boundaries from micro-domains of each RAID in a Fragment-view, our initial idea was to draw a contour map according to its data distribution for each micro-domain and then select a suitable level contour line to limit the boundary. The module of MCI-2kde includes: (i) to construct contour density map for a sub-cluster (micro-domain), MCI-view only takes fragments as data points of a scatter plot despite lines. For each data point, the fragment genomic locus (coordinate of the fragment start position) and the clustered *y*-axis position of the complex it belongs to, are considered as two indices (x and y). (ii) MCI-view then performs 2D KDE over these points within a sub-cluster and returns a result of a matrix of density estimation approximation, consequently the matrix displays the results as contour maps. (iii) Finally, a contour line that encircles the majority of data points of a micro-domain is selected for the microTFY boundary definition.

In this work, MCI-view employs *geom_density_2d* program (*ggplot2* package of R) to call *kde2d* function perform the 2D KDE over a micro-domain fragments and outputs as a contour density map, which generates 10 levels of contours by default. Next, the innermost contour line that enclose at least 60% data points was selected. Finally, the region between the leftmost points to rightmost points of the selected contour line is squared to represent the micro-domain boundary.

Assessment of microTFY boundary defined by MCI-2kde using LabelMe software

LabelMe is an image annotation tool that allows researchers to label images by hand and obtain the annotation information [41]. In this work we drew a square for each micro-domain as its ground truth boundary and marked in a JSON file. For evaluating the potential artificial factors, we executed LabelMe by six different people via a blind approach and obtained microTFY boundaries for further assessment. After that, we wrote a custom script to convert ground truth boundaries information from JSON file to genomic coordinates. Then we compared the boundaries defined by MCI-2kde by calculation intersection over the ground truth by LabelMe for the validation of MCI-2kde boundary definition.

Clustering of the matrix crossing microTFY and TFs

In order to study functions of microTFY, we downloaded FASTQ files of 35 TFs ChIP-seq data of fruit fly from public datasets (Supplementary Table S1). The sequences are mapped to the dm3 reference genome using BWA-ALN and BWA-SAMSE [42], then unique alignments (mapping quality $\geq$ 30) and nonredundant reads retained to hit the genomic regions of the 578 microTFY. Subsequently, RPKM values of individual TFs were projected on each microTFY for a data matrix construction. Next, we performed HCPC clustering [28] on both directions of the data matrix and achieved six microTFY clusters and six TFs clusters, which was shown as a heatmap.

For further dissecting microTFY clusters features, we downloaded ChIP-seq data of the four histone marks (H3K27ac, H3K27me3, H3K4me1, H3K4me3) and RNAPII, as well as RNA-seq data of fruit fly from public datasets (Supplementary Table S1). Those libraries were processed as same as described above, despite of RNA-seq data using STAR [43] as the mapping tool. Next, we hit these signals to the clustered microTFY, and consequently got RPKM values to construct cluster-wised boxplots (the extreme sparse microTFY Cluster-1 was discarded).

Results

Overview of MCIBox

There are two main functions in MCIBox: a browser 'MCI-view' for multiple chromatin interaction data visualization and a framework for micro-domain boundary quantification using 'MCI-2kde' program. MCI-view is developed for the general and comprehensive visualization of multi-way chromatin interactions complexes from the emerging ligation-free based approaches including ChIA-Drop, SPRITE and GAM (Figure 1). MCI-view can be applied to view new aspects of 3D genome topology and microscale chromatin structure, such as the chromatin fiber organization activity during transcription and regulation, the single-molecule chromatin complexes clustered micro-domains in TADs (called microTADs) or the microTFY. MCI-2kde is a 2D KDE algorithm-based unsupervised machine learning method for micro-domain definition automatically.

MCI-view displays contact maps for single-molecule multi-way chromatin complexes

MCI-view is a key component of MCIBox and provides a visualization system for multi-way chromatin interactions. Users can visualize their own experimental data by creating a formatted document that contains information about complexes in each line from the results of multiple chromatin contacts generated by multi-way contacts detected approaches (Figure 1A–D and Supplementary Figure S1; Methods). MCI-view reads the data of the genomic region of interest from the formatted input file for matrix reconstruction (Figure 1E), then integrates Hierarchical clustering strategies to cluster the HD matrix data directly or performs dimensionality reduction methods to reduce the matrix data into LD data for further clustering when the single-molecule complexes data feature is too complicated. This approach was specifically developed for integrative genomics viewer for handling multi-way contact data (Figure 1F and G). MCI-view also can browse accumulated density results for 1D track similar to the coverage of ChIP-seq data and can browse 2D profiles such as loop or domain annotations, side by side simultaneously for conveniently compare to pairwise contacts (Figure 1H and Supplementary Figure S1A–C).

For single-molecule chromatin interactions, the interface of MCI-view supplies a function to display the data by specifying the interested single or multiple genomic coordinates, or the interested gene name and an additional function to filter out the uninterested data. Currently, MCI-view includes six modules for displaying multi-way contact results (Supplementary Figure S1): (i) **Cluster-view** performs binning of clustering tracks for browsing the clustered microscale domains of multiplex chromatin interaction complexes (Supplementary Figure S1D and G). (ii) **Fragment-view** directly visualizes the original (unbinned) fragments for the multi-way contacts representation (Supplementary Figure S1E and H). (iii) **Target loci view** presents the specific region similarly to DNA-FISH probe associated genomic loci, selected by clicking on the interested regions (Figure 1K and Supplementary Figure S1I). (iv) **Chromatin organization pattern view** is used for observing the multiple chromatin interactions at TF binding motifs such as CTCF motifs (Figure 1J and Supplementary Figure S1J). (v) **Transcription pattern view** is used for discovering the multiple chromatin interactions profiling during the process of gene transcription in order to explore the extrusion process (Figure 1I and Supplementary Figure S1K). (vi) **Transcription regulation pattern view** can be applied for observation of the multi-way chromatin interactions at super-enhancers region [16], thereby allowing users to identify the composition of super-enhancers for gene regulation.

MCI-view uncovers super-enhancers interacting in multi-way contacts

To explore the contact detail of super-enhancers with the working model as a whole unit or by different composed enhancers to regulate gene promoter in individual nuclei (Figure 2A), we used MCI-view to display the interacting profiling of super-enhancers from mouse ESC GAM data in Figure 2B (same genomic region in '*Fig. 5a*' of the GAM paper [16]). The 2D heatmap shows the pairwise contacts from GAM results with square highlight of enhancer contact regions. Cluster-view shows the different composed enhancers in the multiple chromatin complexes, along with the composition of micro-topologically associated domains (microTADs) that are accumulated into the TAD (Figure 2B). In addition, Fragment-view displays detailed original fragments clustering profiling for the super-enhancers located in microTADs (Figure 2C). This particular genomic region with three enhancers showed that 53.7% of multiplex chromatin complexes involve one enhancer, 37% involve two enhancers and 9.3% of them involve three enhancers. Thus, the composition of enhancers can be variable in the super-enhancer region for gene-specific regulation in individual cells (Figure 2D).

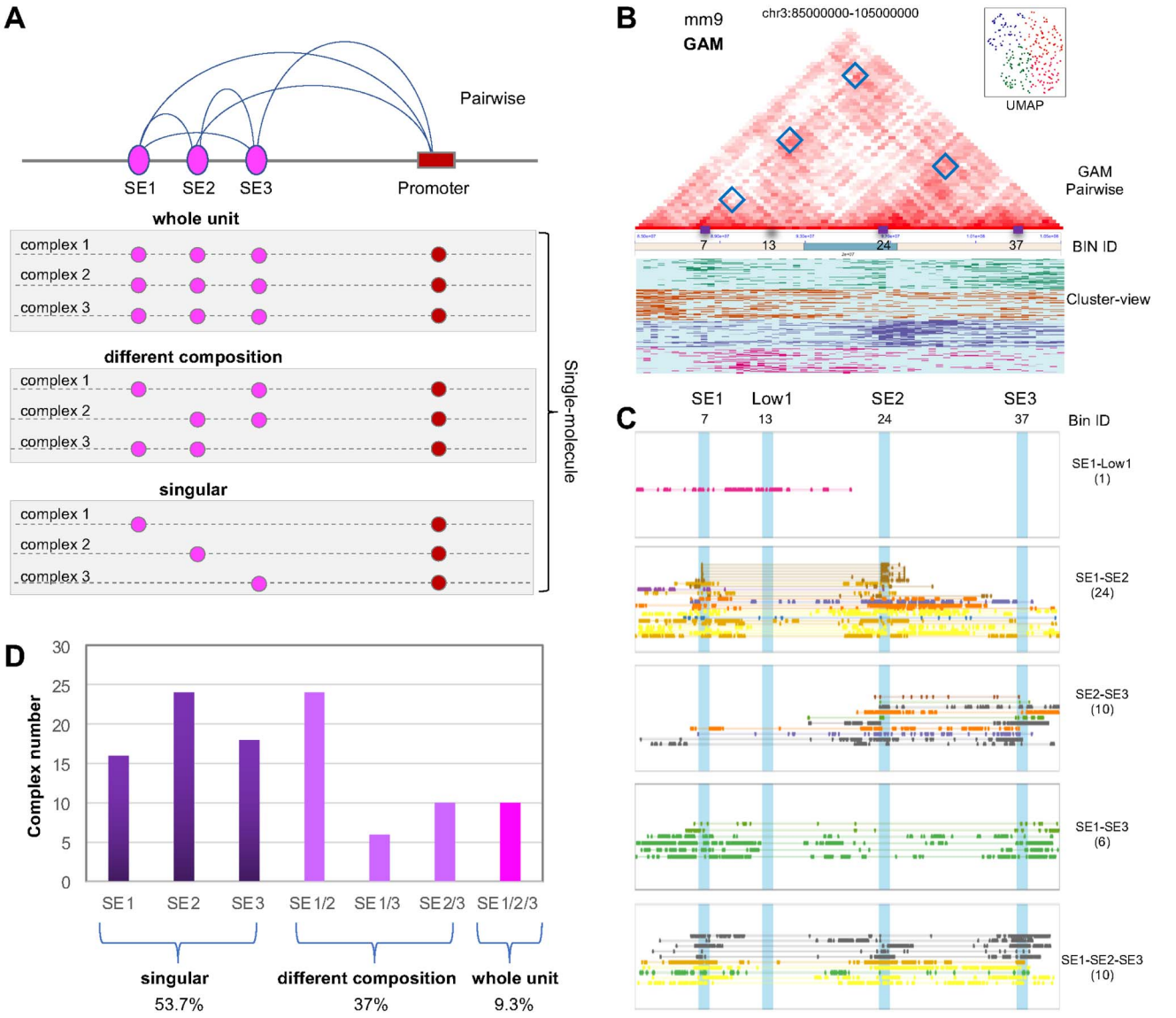

Figure 2. MCI-view displays super-enhancers working model. (**A**) Diagram of the potential working model for Super-enhancers (SEs) on gene regulation, 'whole unit' working model presents all the enhancers in a super-enhancer region are tethered together to regulate target genes in different cells, 'different composition' working model indicates different composed of enhancers co-regulate in different cells, 'singular' working model shows only one enhancer regulates target genes in different cells. (**B**) MCI-view displays the pairwise contacts of the genomic region (chr3:85000000-105000000) from mouse GAM data, from which we could find the selected enhancer points annotated with square shown in the heatmap ('*Fig. 5a*' from the GAM paper [16]). Besides this, the Cluster-view function in MCI-view shows that there are several micro-domains, which indicates different chromatin interaction clustering within this genomic region. Inset scatter plot presents the clustering result. (**C**) 'Transcription regulation pattern' in MCI-view to display the contact profiling of these super-enhancer regions. Number of '7, 13, 24 and 37' presents the Bin ID in this region and corresponding to 'SE1, Low1, SE2 and SE3' in the GAM paper [16]. SE1, SE2 and SE3 present super-enhancer 1, 2 and 3 loci. Low1 presents a control region not involved in the contacts among these super-enhancers. SE1-Low1 presents the interaction between super-enhancer SE1 and Low1 region, the number inside brackets shows the detected multi-way contacts, so and forth. (**D**) Histogram shows the percentage of chromatin complexes for composed enhancers among SE1, SE2 and SE3.

MCI-view directly visualizes asymmetric loop-extrusion process

Chromosomes are organized as chromatin loops to promote the interactions between enhancer–promoter, allowing for the long-range regulation of gene expression. Loops were hypothesized to form by 'loop extrusion' and visualized by real-time imaging in vitro [44]. Our RNAPII ChIA-drop method captured multiplex interactions on *Drosophila* S2 cell line data presenting the asymmetric chromatin extrusion during gene transcription genome wide. Here we applied human GM12878 SPRITE data to explore and display the loop extrusion associated with CTCF, which has not been analyzed in SPRITE paper since no published tools for visualization at that time.

The previously published SPRITE dataset [17] was converted to the formatted document for MCI-view visualization. As shown in Supplementary Figure S2A, the accumulated 2D heatmap shows pairwise contacts at the selected genomic region, whereas the 1D coverage shows variable density at this region corresponding to the higher-order structure of a domain on 2D heatmap. The key is that the Cluster-view module of MCI-view unfolds the layers for the detailed composition of multiple chromatin complexes in this region that were clustered into nine micro-domains.

Fragment-view directly displays the original fragments information of these clustered chromatin complexes. Due to the MCI-view function for selecting the anchor of interests, we can clearly visualize the multi-way contacts by Cluster-view. In the case of this targeted region, the view agrees well with the previously published results (in '*Figure S2D*' of the SPRITE paper [17]) and can further observe the original contact loci of the chromatin complexes by Fragment-view in MCI-view.

Moreover, we can observe the chromatin extrusion pattern from the multi-way contacts-associated CTCF binding motifs via the module of 'chromatin organization pattern view' in MCI-view (Supplementary Figure S2B–D and Figure 3). Within this genomic region (Figure 3A–D), there are 913 chromatin complexes that contained at least one CTCF motifs at left or right sides, and there are 466 of them that covered the left-side CTCF motifs, 364 of them that covered the right-side CTCF motif, only 83 (9%) of them covered both sides. These results indicate that CTCF-associated chromatin complexes exhibit one-sided DNA loop extrusion profiling, where CTCF complexes stop and tether at DNA loci with CTCF binding motifs, then reel the other side of DNA to close gradually and stop until meet another CTCF binding site with 'convergent CTCF motifs', these complexes are not cross over the motifs. Generally, the CTCF motifs at the boundary of CCDs (CTCF-mediated chromatin domains) are in a convergent orientation [6]. We quantified the single-molecule chromatin complexes within a CCD that overlapped with the convergent CTCF sites on the boundary for these previously defined 2266 of CCDs and found that the percentage of multi-way contacts covered one of the motifs is >90% (Figure 3D, inset boxplot).

Our observation of loop extraction activity in MCI-view is consistent with the findings of real-time imaging of asymmetric DNA loop extrusion that condensin anchors onto DNA and reels it in from only one side [44], and the findings of live-cell imaging of loops covered both sides are rare [45]. Additionally, RNAPII ChIA-Drop results provide a promoter-centred extrusion model during gene transcription (Supplementary Figure S2E) [18]. We note that the putative higher-order structure domain of 2D heatmap from Hi-C data supports the notion that chromatin loops extrude asymmetrically. Otherwise, if they extrude symmetrically, they would exhibit a secondary diagonal overlaid on the main diagonal as a '+' shaped pattern (Figure 3E).

MCI-view displays micro-domains with clustered single-molecule chromatin complexes

The proposed model of 'transcription factories' that RNA polymerases are immobilized and concentrated within discrete foci in nuclei for multiple genes transcription and regulation [15]. Data from imaging to genomics offer the evidence for RNAPII foci and RNAPII-mediated promoter–enhancer clusters, i.e. 'transcription factories', participate in regulation of gene expression (Figure 4A) [3, 46, 47]. However, the corresponding genomic coordinates of individual RNAPII foci has yet to be studied, and RNAPII clusters were accumulated from bulk cells, which cannot reveal the precise composition of transcription factories.

The advanced ligation-free chromosome conformation capture methods RNAPII ChIA-Drop could decompose the transcription factories into single-molecule precision. Applying MCI-view to display the single-molecule chromatin complexes from RNAPII ChIA-Drop data with a dimensionality reduction algorithm followed by a clustering algorithm, when combine UMAP and Hk-means algorithm we can obtain high silhouette score for high-quality clustering, based on that we found that the RNAPII-mediated chromatin interaction domains (RAIDs) can be organized by several different micro-domains, named microTFY (Figures 1G and H, 4B and C and Supplementary Figure S3A–C). This observation indicates that the micro-domains organized by a certain similarity from single-molecule chromatin contacts could potentially offer a new pathway for elucidating gene regulation (Supplementary Figure S3C).

MCI-2kde automatically determines microTFY boundaries via density estimation

To automatically quantify the boundary of microTFY, we constructed contour density map on the microTFY clusters from MCI-view by performing a 2D KDE algorithm [40]. Within one microTFY, if the points are of the same range of density distribution probability, they will be enclosed by the same closed curve (contour line). Based on the density contour map of each microTFY, we determined the boundary of a microTFY by selecting the contour line with concentrated coverage of the fragments of chromatin complexes in Fragment-view (Figure 4D and E; Methods).

Using this method, we identified 578 microTFY using RNAPII ChIA-Drop data through fixed genomic regions from the previously defined 126 RAIDs with accumulated RNAPII ChIA-PET data. For validation of the microTFY boundary identification, we then enter the image of microTFY from MCI-view to LabelMe, a database and web-based tool for image annotation [41], for recognizing the objects of microTFY and labeling it along the boundary of microTFY (Figure 4F). Based on LabelMe, we can obtain the ground truth boundary of microTFY by manually importing images one by one.

When we compared the boundaries of microTFY between the identification of density contour map via MCI-2kde program and that of LabelMe upon the image of microTFY from MCI-view directly, we found the intersection over the ground truth from LabelMe executed by different people is about 90% when contour line setting with 60% coverage of the chromatin complexes. These indicate the approaches allow us to determine the microTFY (Figure 4G and Supplementary Figure S4).

In summary, we developed a framework for helping scientists to automatically define the boundary of microTFY. This framework was integrated in MCIBox and can be easily extended to the detection of other types of micro-domains generated from 3D genome mapping technologies for exploring new aspects of higher-order chromatin structure.

Characterization of micro-domains from single-molecule chromatin complexes

The genomic feature of microTFY shown in the density plot with a peak of concentrated gene number at 6 in individual microTFY, with genomic size at a peak of 30-kb length (Figure 5A). The previously defined RAIDs by pairwise data were composed with microTFY by single-molecule chromatin complexes. We found that the genes from different phases of a cell cycle are distributed at different microTFY within a RAID, for example, the subG1 phase-associated *CG17209* is located in the microTFY on the left side and the G1 phase-related *Myb* is located in another microTFY on the right side (Figure 5B) [48]. Additionally, the overG2 phase associated *scra*, *sax* and G1 phase-associated *tor* are all located in individual microTFY that within in a RAID (Figure 5C) [48]. Overall, the phase-specific genes of a cell cycle could be identified in different microTFY via single-molecule chromatin contacts while before being regarded as in a same RAID by pairwise chromatin interactions (Figure 5D).

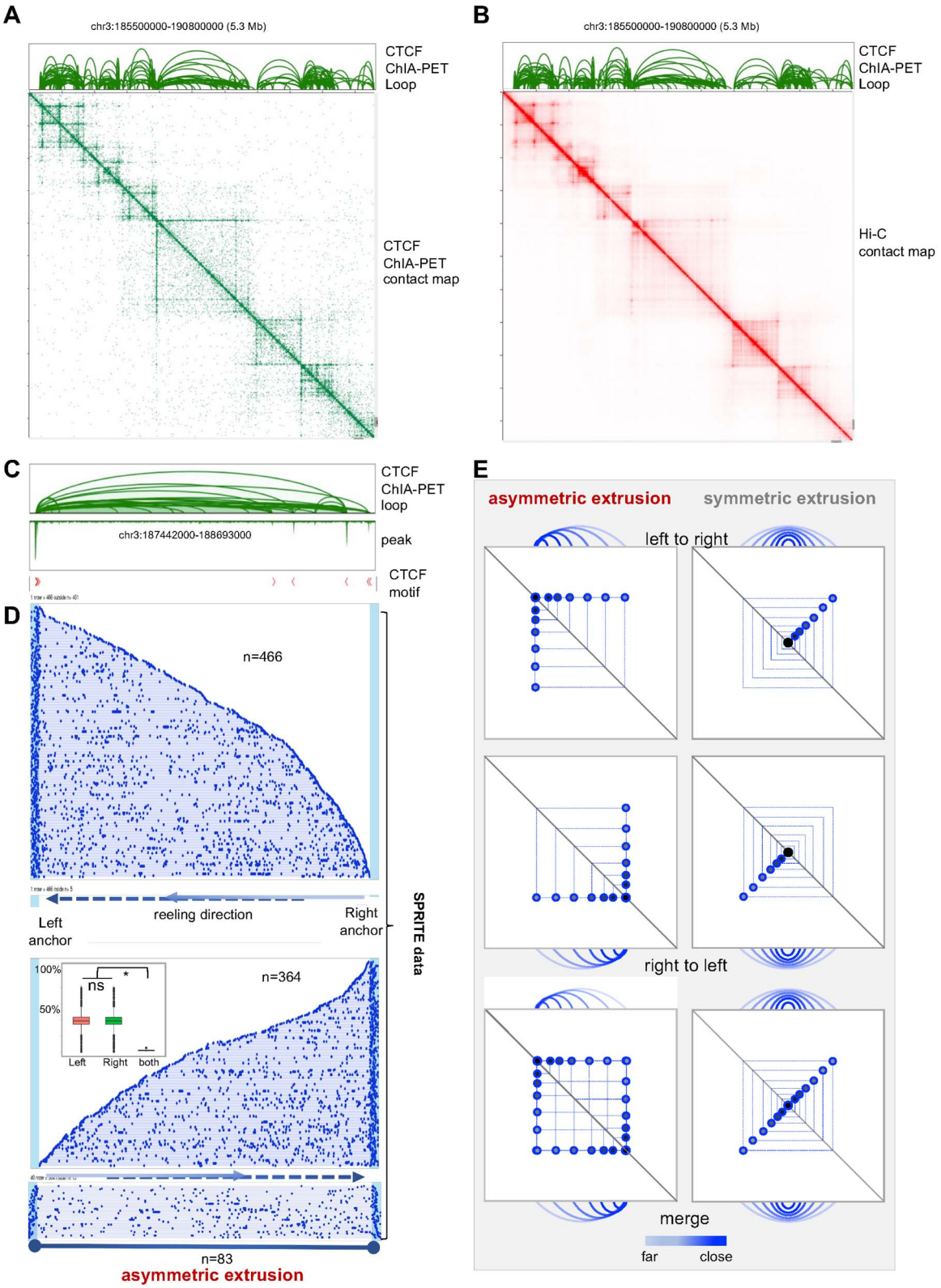

Figure 3. MCI-view displays chromatin loops extrusion. (**A**–**B**) BASIC Browser displays the pairwise contact loops of the genomic region (chr3:185500000-190800000) from CTCF ChIA-PET data on human GM12878 cells, along with 2D heatmap from CTCF ChIA-PET data (green) and Hi-C data (red). (**C**) BASIC Browser displays the pairwise contact loops and binding peaks of the zoomed-in genomic region (chr3:187442000-188693000) from CTCF ChIA-PET data on human GM12878 cells. Arrow with red color indicates CTCF binding motifs. (**D**) MCI-view displays the single-molecule chromatin complexes of the genomic region (chr3:187442000-188693000) from SPRITE data on human GM12878 cells [17]. MCI-view displays SPRITE multiplex chromatin complexes covered the left anchor contained CTCF motifs with cyan color, multiplex chromatin complexes were shown as Fragment-view aligned by complexes span length in the local region, followed by that of right anchor and both anchors. Line with arrows present chromatin loops reeling direction, line with solid circle indicates chromatin loops stop reeling that means no complexes cross over the boundary of a CCD or TAD, corresponding to CTCF binding motifs with convergent orientation generally, the number (*n*) of chromatin complexes overlapped with CTCF motifs is shown, which

To infer genome wide the characteristics of microTFY, we used 35 TF and histone marker ChIP-seq datasets in *Drosophila* S2 cells from the modENCODE project (Supplementary Table S1). Then, we investigated the association between these protein factors and the 578 microTFY by HCPC clustering (Figure 5E; Methods). These 35 TFs are functionally clustered into different types (TF cluster in x-axis; Figure 5E), such as: (TF-iii) Dosage compensation-related factors including Msl1, Msl3 and Mle. (TF-iv) Transcription-associated factors including Rpb1, Ash1, Dsx, Smt3, Fs(1)h and Lpt. (TF-v) Chromosome organization-related factors including Rrp40, Ago2, Psc, Trr, Ms(3)K81 and CTCF. (TF-vi) Chromosome organization regulation-associated factors including CP190, Ice1 and Yki, etc. We found these six clusters of microTFY are involved in different types of transcription factors (microTFY cluster in y-axis; Figure 5E). For instance, Cluster-3 and Cluster-4 consisted of 47 and 100 microTFY that were bound by most of the 35 TFs with different density, respectively; Cluster-5 included 223 microTFY mostly bound by factors with transcription-associated functions and dosage compensation-related functions and Cluster-6 with 197 microTFY showed mostly week binding by most of the factors.

To further characterize the epigenomic feature and transcriptional activity of the microTFY clusters from RNAPII ChIA-Drop data (Figure 5E), we aligned them with histone ChIP-seq and RNA-seq data. Overall, we observed that the genes and transcription activity had high variability in these microTFY clusters (Figure 5F). As expected, the Cluster-6 microTFY exhibited low signals of H3K27ac (active histone marker), H3K4me1 (enhancer marker), H3K4me3 (promoter marker), RNAPII (transcription marker) and RNA-seq, as well as high signal of inactive histone marker H3K27me3. By contrast, Cluster-2 and Cluster-3 microTFY showed strong signals of H3K27ac, H3K4me1, H3K4me3, RNAPII and RNA-seq. This observation suggests that each cluster of microTFY is involved in different types of transcription activity and histone modification.

Taken together, these results demonstrate the distinct ability and advantages of MCIBox to conveniently and effectively browse 3D genome features for the single-molecule chromatin complexes and identify the boundaries of microTFY automatically. The boundary determination of microTFY allows characterization of cell-cycle-specific or cell-type-specific gene regulation by multiplex chromatin interactions with single-molecule precision. MCIBox is an invaluable tool in making biological discoveries from many aspects for multi-way contacts data.

Discussion

In this work, we developed MCIBox for visualizing multi-way chromatin interactions and automatically identifying micro-domain boundaries. MCIBox allows for comprehensive browsing of the single-molecule multi-way chromatin interaction data generated by the cutting-edge ligation-free 3D genome mapping technologies, such as SPRITE and ChIA-Drop. MCIBox includes an efficient MCI-view module for displaying 3D genome features and chromatin structures. MCI-view can unveil microscale structure in 3D visual panel of Cluster-view and Fragment-view and can also facilitate exploration of the extrusion model of chromatin organization by CTCF motif-based for chromatin organization pattern view and promoter-based for transcription pattern view. Additionally, we developed frameworks MCI-2kde in MCIBox for automatically definition of the boundary of micro-domains, revealing genomic and epigenomic features of microTFY in strong association with gene regulation.

MCI-view allows for the application of a selected clustering method (Hierarchical clustering, Hk-means etc.) with, or without a dimensionality reduction algorithm (UMAP, TSNE etc.) in the preceding clustering, in order to display the interactive data of multi-fragment from the ligation-free based new generation 3D genome techniques (ChIA-Drop, SPRITE or GAM) in forms of multi-type of tracks (Cluster-view, Fragment-view, etc.) for regions of interest in a genome browser. In this study, we introduce silhouette score to evaluate the quality of clusters. Users can select a suitable algorithm for their data clustering by score rating. For each parameter used in the clustering algorithms, MCIBox starts from its default value defined in the function implementing the corresponding algorithm, but the user can also adjust them freely for better clustering results.

Currently, there exist more than 20 approaches for the boundary definition for chromatin structures such as TADs, yet most of them are based on profiles calculated from 2D heatmaps or epigenomic information, such as linear scores, statistical models of the interaction distributions, clustering and network features [49]. The TAD boundaries that called by statistics often show some shifts as compared that of observation by browser directly.

To overcome these challenges, we adopted the strategy of geometric topographic map in the MCI-2kde module, where we drew a density contour map by 2D KDE in individual sub-cluster from Fragment-view and successively derived the boundary of a microTFY through contour line selection. From this density contour map, we obtained 578 microTFY automatically.

As comparison of intersection region of microTFY among the boundary identified approaches LabelMe and MCI-2kde, we found the percentage of intersection over the ground truth is about 90% when the contour line covered chromatin complexes occupied 60%. Thus, the users could according to their requirement to adopt the method for boundary determination. For MCI-2kde module, the default cutoff of contour line is 60% to obtain the corresponded area covered chromatin complexes, we can increase the cutoff value by compromised for covered more chromatin complexes while with sparse distribution.

Finally, MCIBox can potentially be extended to analyze single-cell assays for higher-order chromatin structures in the future by input data of multi-way contacts in single cells such as scSPRITE data. MCIBox, in its current form, is a convenient tool kit for single-molecule chromatin complexes analysis, which not only systemically offers a visual interface for exploring the pattern of chromatin organization, transcription and regulation in single-molecule level, but also provides a platform for characterizing the micro-domains detected automatically from the clustered multi-way chromatin contacts. MCIBox is capable of potentially extending to distinguish the chromatin organization activity of

can be directly observed in MCI-view browser. The inset boxplot presents globally quantify these architectural stripes in 2266 of CCDs, y-axis shows the percentage of chromatin complexes occupied the boundary of CCDs, x-axis shows the boundary of CCD including left-, right- and both sides. The boxplot with lines as medians and whiskers at ±1.5 interquartile range; Two-sided Wilcoxon test, * P-value was $2.2x10^{-16}$. (**E**) Diagram of asymmetric and symmetric chromatin loops extrusion is shown. Arch diagram showing positions of one-sided or two-sided CTCF complexes translocating along chromatin DNA (gray) and progressively growing loops at different times indicated by color from blue (close) to transparent blue (far). The drawing 2D contacts corresponding the asymmetric and symmetric loops extrusion is shown under the arch diagram. Chromatin loops translocating along chromatin DNA from left to right (Top), right to left (Middle), merge of both (Bottom), are shown.

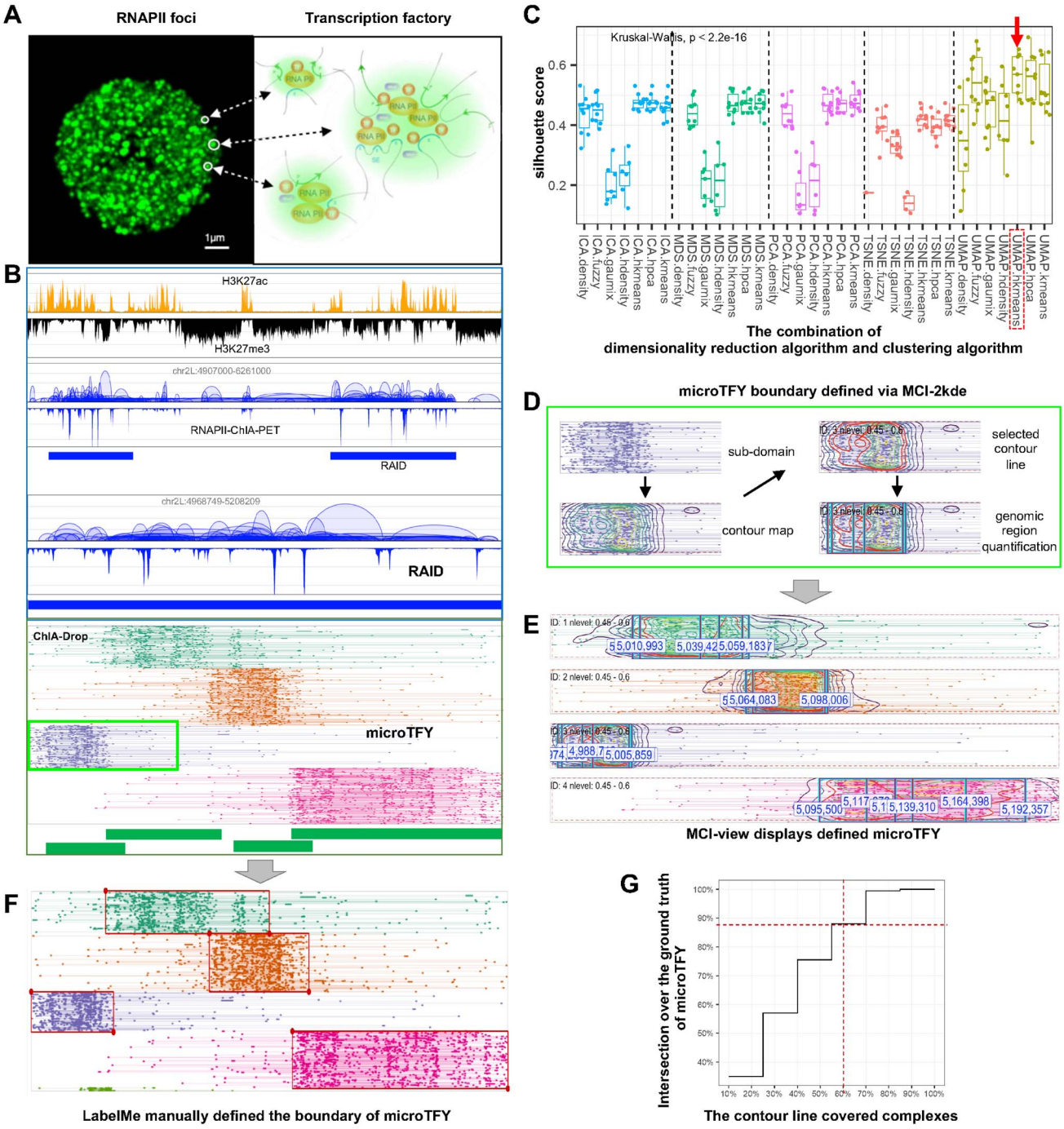

Figure 4. Definition of microTFY boundary by MCI-2kde. (**A**) RNAPII immunostaining results show different sizes of RNAPII foci, supporting the concept of transcription factories. (**B**) BASIC Browser displays pairwise contacts of RNAPII ChIA-PET data and MCI-view displays multi-way contacts of RNAPII ChIA-Drop. RNAPII ChIA-PET data show interaction loops and RNAPII coverage density, ChIP-seq data of H3K27ac and H3K27me3 shows the active and inactive regions, the blue bar presents transcription factories that were defined from RNAPII ChIA-PET data. Each transcription factory is composed by several microTFY (green bar) detected from RNAPII ChIA-Drop results. (**C**) Silhouette score for the combination of dimensionality reduction algorithm and clustering algorithm. Red arrow shows the high silhouette score for UMAP and Hk-means. The boxplot with lines as medians and whiskers at ±1.5 interquartile range; Two-sided Kruskal–Wallis test, *P*-value was shown. (**D**) Definition of microTFY boundary via MCI-2kde. Density contour map was constructed by fragment distribution probability in each individual clusters shown in Fragment-view, then demonstrated the innermost contour line that covered at least 60% data points, and finally quantified their genomic coordinates by the extremums along the horizontal direction. (**E**) MCI-view displays microTFY boundary defined by MCI-2kde. The minimal coordinate and maximal coordinate of the square represents the boundary of the corresponded microTFY. (**F**) The identification of microTFY boundary via LabelMe software. Red squares indicate the defined boundaries for individual microTFY. (**G**) Empirical cumulative distribution function curve shows the intersection over the ground truth of microTFY. Red dashline highlighted when the contour line covered chromatin complexes occupied 60% the intersection over the ground truth is almost 90%.

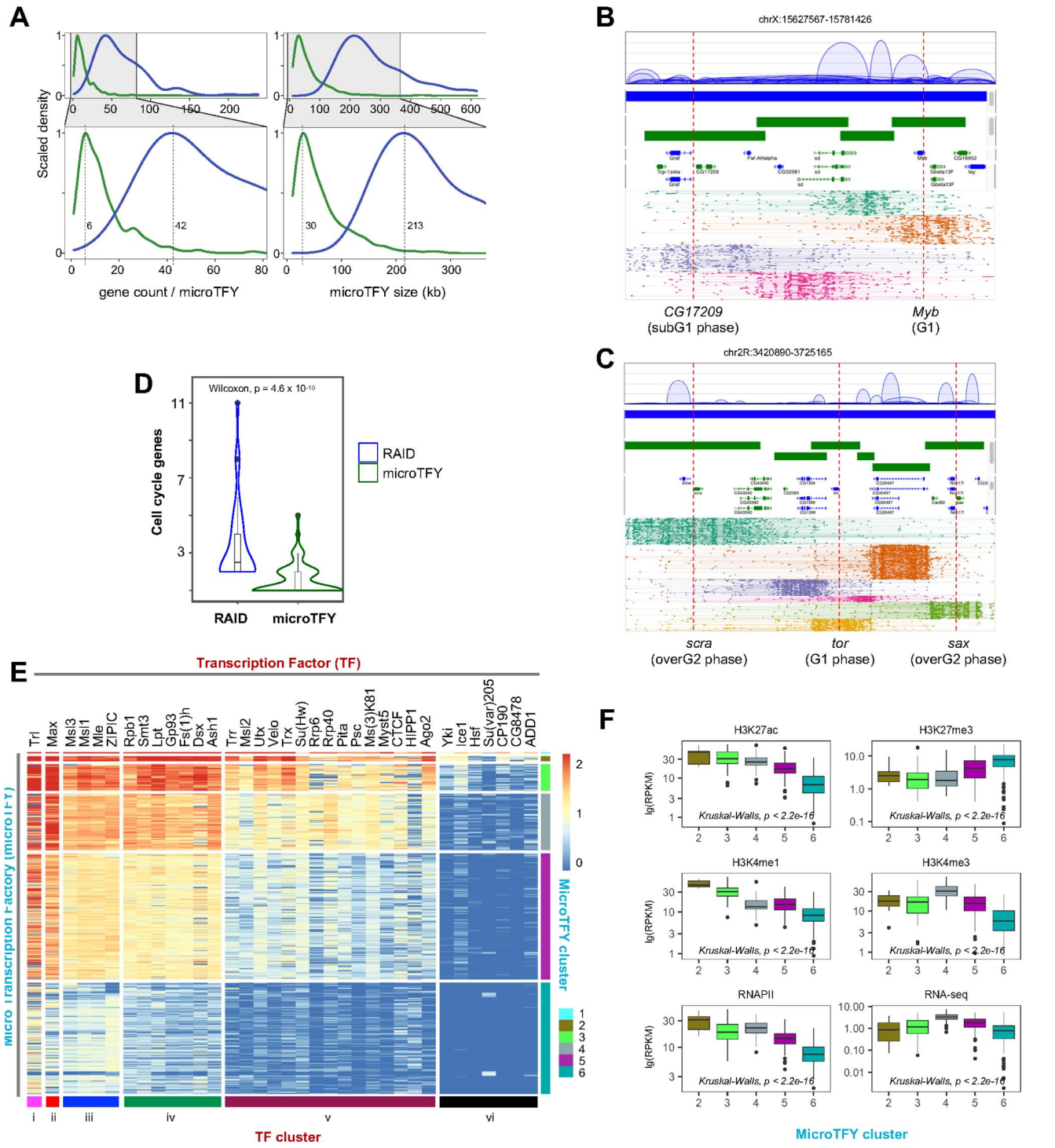

Figure 5. Characterization of microTFY from RNAPII ChIA-Drop data. (**A**) Gene count distribution. The scaled density of gene count for the defined 578 of microTFY are plotted, with a peak at 6 of genes for microTFY (green) and 42 of genes for RAIDs (blue), and the scaled density of microTFY and RAIDs size were plotted, with a peak at 30 and 213 kb, respectively. (**B**–**C**) RAIDs (blue bar) present multiplex chromatin interactions from RNAPII ChIA-PET data are composed with several microTFY when display with RNAPII ChIA-Drop data (green bar). Panel B shows a RAID that was composed by four microTFY, left one contained cell cycle subG1 phase-associated gene *CG17209*, right one contained cell cycle G1 phase-related gene *Myb*. Panel C shows a RAID that was composed by six microTFY, left one contained cell cycle overG2 phase-associated gene *scra*, middle one contained cell cycle G1 phase-associated gene *tor*, right one contained cell cycle overG2 phase-related gene *sax*. (**D**)The violin plots of cell cycle gene counts for RAID and microTFY was shown. The boxplot with lines as medians and whiskers at ±1.5 interquartile range; Two-sided Wilcoxon test, *P*-value was shown. (**E**) We use 35 TF and histone marker ChIP-seq datasets in *Drosophila* S2 cells from the modENCODE project to cluster the 578 microTFY into six groups by HCPC clustering. *y*-axis shows microTFY and *x*-axis presents TFs. Bar with color indicates different clusters. (**F**) Epigenetic feature for microTFY. The boxplot of ChIP-seq data for H3K27ac (active marker), H3K4me1(enhancer marker), H3K27me3 (inactive marker), H3K4me3 (promoter marker), RNAPII (promoter and enhancer marker) and RNA-seq data for gene expression were plotted respectively (Cluster-1 is not shown here as its data is too sparse). The boxplot with lines as medians and whiskers at ±1.5 interquartile range; Two-sided Kruskal–Walls test, *P*-value was shown.

cell cycle specificity and even cell type specificity using single-molecule chromatin contacts, to identify the chromatin extrusion model and super-enhancers regulation model.

Key Points

- We describe a new toolkit MCIBox, which includes a visual tool MCI-view for multi-way chromatin interactions visualization and a platform for single-molecule chromatin complexes clustered micro-domains determination.
- MCIBox is based on various clustering algorithms integrated with dimensionality reduction methods that can display multiplex chromatin interactions at single-molecule level.
- We demonstrate MCIBox ability to explore chromatin extrusion patterns and super-enhancers regulation modes in transcription, and to identify single-molecule chromatin complexes that cluster into micro-domains.
- MCIBox incorporates machine learning algorithms to identify micro-domains boundaries automatically. These micro-domains were stratified with distinctive signatures of transcription activity and contained different cell-cycle-associated genes, respectively.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*

Data Availability

The source code of MCIBox can be accessed at the public GitHub repository (https://github.com/ZhengmzLab/MCIBox). The publicly available datasets used in this study are as follow: ChIA-Drop data from Zheng *et al.* [18] (GEO:GSE109355); SPRITE data from Quinodoz *et al.* [17] (GEO:GSE114242); GAM data from Beagrie *et al.* [16] (GEO:GSE64881) and the available datasets for ChIP-seq and RNA-seq data are shown in Supplementary Table S1.

Funding

National Natural Science Foundation of China (32170644); Shenzhen Innovation Committee of Science and Technology (ZDSYS20200811144002008). M.J.F. is supported by the National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centres of Excellence initiative and by a Ministry of Education Tier II grant awarded to M.J.F. (T2EP30120-0020). D.P. is co-supported by Warsaw University of Technology within the Excellence Initiative-Research University (IDUB) programme, co-supported by Polish National Science Centre (2019/35/O/ST6/02484 and 2020/37/B/NZ2/03757) and EU-funded the Marie Sklodowska-Curie action (MSCA) Innovative Training Network.

Acknowledgements

The authors are grateful to Dr Yijun Ruan for suggestions on manuscript organization and Dr Daniel Capurso for suggestions on the improvement of MCI-view functions. The authors thank Dr Minji Kim for editing and polishing this manuscript.

References

1. Dekker J. Gene regulation in the third dimension. *Science* 2008;**319**(5871):1793–4.
2. Fullwood MJ, Liu MH, Pan YF, *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* 2009;**462**(7269):58–64.
3. Li G, Ruan X, Auerbach RK, *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 2012;**148**(1):84–98.
4. Lieberman-Aiden E, van Berkum NL, Williams L, *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;**326**(5950):289–93.
5. Rao SSP, Huntley MH, Durand NC, *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**(7):1665–80.
6. Tang Z, Luo OJ, Li X, *et al.* CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* 2015;**163**(7):1611–27.
7. Durand NC, Robinson JT, Shamim MS, *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* 2016;**3**(1):99–101.
8. Kerpedjiev P, Abdennur N, Lekschas F, *et al.* HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol* 2018;**19**(125):1–12.
9. Akdemir KC, Chin L. HiCPlotter integrates genomic data with interaction matrices. *Genome Biol* 2015;**16**(198):1–8.
10. Wang Y, Song F, Zhang B, *et al.* The 3D genome browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol* 2018;**19**(151):1–12.
11. Zhou X, Lowdon RF, Li D, *et al.* Exploring long-range genome interactions using the WashU epigenome browser. *Nat Methods* 2013;**10**(5):375–6.
12. Tang B, Li F, Li J, *et al.* Delta: a new web-based 3D genome visualization and analysis platform. *Bioinformatics* 2018;**34**(8):1409–10.
13. Lee B, Wang J, Cai L, *et al.* ChIA-PIPE: a fully automated pipeline for comprehensive ChIA-PET data analysis and visualization. *Sci Adv* 2020;**6**(28):eaay2078.
14. Zhu X, Yang Z, Wang Y, *et al.* Nucleome browser: an integrative and multimodal data navigation platform for 4D nucleome. *Nat Methods* 2022;1–3.
15. Cook PR. The organization of replication and transcription. *Science* 1999;**284**(5421):1790–5.
16. Beagrie RA, Scialdone A, Schueler M, *et al.* Complex multi-enhancer contacts captured by genome architecture mapping (GAM). *Nature* 2017;**543**(7646):519–24.
17. Quinodoz SA, Ollikainen N, Tabak B, *et al.* Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* 2018;**174**(3):744–757.e24.
18. Zheng M, Tian SZ, Capurso D, *et al.* Multiplex chromatin interactions with single-molecule precision. *Nature* 2019;**566**(7745):558–62.
19. Kempfer R, Pombo A. Methods for mapping 3D chromosome architecture. *Nat Rev Genet* 2020;**21**(4):207–26.
20. Phillips-Cremins JE, Sauria MEG, Sanyal A, *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* 2013;**153**(6):1281–95.

21. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 1987;**20**:53–65.
22. Kaufman L, Rousseeuw J. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Mathematical Statistics. Wiley, Hoboken, NJ, 2005.
23. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. *WIREs Data Mining Knowl Discov* 2012;**2**(1):86–97.
24. Wu X, Kumar V, Quinlan JR, *et al.* Top 10 algorithms in data mining. *Knowl Inform Syst* 2008;**14**(1):1–37.
25. Schubert E, Sander J, Ester M, *et al.* DBSCAN revisited, revisited: why and How you should (still) use DBSCAN. *ACM Trans Database Syst* 2017;**42**(3):1–21.
26. Scrucca L, Michael Fop T, Murphy B, *et al.* mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *R J* 2016;**8**(1):289–317.
27. Varada Rajkumar K, Yesubabu A, Subrahmanyam K. Fuzzy clustering and fuzzy c-means partition cluster analysis and validation studies on a subset of citescore dataset. *Int J Electrical Comput Eng (IJECE)* 2019;**9**(4):2760–70.
28. Husson F, Josse J, Pages J. Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data? *Agrocampus* 2010;1–17.
29. Tian-Shi X, Chiang H-D, Liu G-Y, *et al.* Hierarchical K-means method for clustering large-scale advanced metering infrastructure data. *IEEE Trans Power Delivery* 2017;**32**(2):609–16.
30. Campello RJGB, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In Pei J, Tseng VS, Cao L, Motoda H, Guandong X, editors, *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, pages 160–72, Berlin, Heidelberg, 2013. Springer.
31. Feng C, Liu S, Zhang H, *et al.* Dimension reduction and clustering models for single-cell rna sequencing data: a comparative study. *Int J Mol Sci* 2020;**21**(6):2181–202.
32. Xiang R, Wang W, Yang L, *et al.* A comparison for dimensionality reduction methods of single-cell RNA-seq data. *Front Genet* 2021;**12**:646936.
33. Zhang T, Yang B. Big data dimension reduction using PCA. In: *2016 IEEE International Conference on Smart Cloud (SmartCloud)*. New York, NY: IEEE, 2016, 152–7.
34. Liebermeister W. Linear modes of gene expression determined by independent component analysis. *Bioinformatics* 2002;**18**(1):51–60.
35. Buja A, Swayne DF, Littman ML, *et al.* Data visualization with multidimensional scaling. *J Comput Graph Stat* 2008;**17**(2):444–72.
36. Linderman GC, Rachh M, Hoskins JG, *et al.* Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat Methods* 2019;**16**(3):243–5.
37. Becht E, McInnes L, Healy J, *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 2019;**37**(1):38–44.
38. Charte D, Charte F, García S, *et al.* A practical tutorial on autoencoders for nonlinear feature fusion: taxonomy, models, software and guidelines. *Inform Fusion* 2018;**44**:78–96.
39. Moon KR, van Dijk D, Wang Z, *et al.* PHATE: a dimensionality reduction method for visualizing trajectory structures in high-dimensional biological data. bioRxiv. 2017;120378.
40. Weglarczyk S. Kernel density estimation and its application. *ITM Web Conf* 2018;**23**:00037.
41. Russell BC, Torralba A, Murphy KP, *et al.* LabelMe: a database and web-based tool for image annotation. *Int J Comput Vis* 2008;**77**(1):157–73.
42. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**(14):1754–60.
43. Dobin A, Davis CA, Schlesinger F, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**(1):15–21.
44. Ganji M, Shaltiel IA, Bisht S, *et al.* Real-time imaging of DNA loop extrusion by condensin. *Science* 2018;**360**:102–5.
45. Gabriele M, Brandão HB, Grosse-Holz S, *et al.* Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging. *Science* 2022;**376**(6592):496–501.
46. Chen X, Wei M, Zheng MM, *et al.* Study of RNA polymerase ii clustering inside live-cell nuclei using Bayesian nanoscopy. *ACS Nano* 2016;**10**(2):2447–54.
47. Wang P, Tang Z, Lee B, *et al.* Chromatin topology reorganization and transcription repression by PML-RARα in acute promyeloid leukemia. *Genome Biol* 2020;**21**(110):1–21.
48. Björklund M, Taipale M, Varjosalo M, *et al.* Identification of pathways regulating cell size and cell-cycle progression by RNAi. *Nature* 2006;**439**(7079):1009–13.
49. Zufferey M, Tavernari D, Oricchio E, *et al.* Comparison of computational methods for the identification of topologically associating domains. *Genome Biol* 2018;**19**(217):1–18.