Nucleic Acids Research, 2021 **1**
https://doi.org/10.1093/nar/gkab1233

cLoops2: a full-stack comprehensive analytical tool for chromatin interactions

Yaqiang Cao †, Shuai Liu†, Gang Ren†, Qingsong Tang and Keji Zhao *

Laboratory of Epigenome Biology, Systems Biology Center, National Heart, Lung and Blood Institute, NIH, Bethesda, MD 20892, USA

Received July 20, 2021; Revised November 18, 2021; Editorial Decision November 29, 2021; Accepted December 02, 2021

ABSTRACT

Investigating chromatin interactions between regulatory regions such as enhancer and promoter elements is vital for understanding the regulation of gene expression. Compared to Hi-C and its variants, the emerging 3D mapping technologies focusing on enriched signals, such as TrAC-looping, reduce the sequencing cost and provide higher interaction resolution for cis-regulatory elements. A robust pipeline is needed for the comprehensive interpretation of these data, especially for loop-centric analysis. Therefore, we have developed a new versatile tool named cLoops2 for the full-stack analysis of these 3D chromatin interaction data. cLoops2 consists of core modules for peak-calling, loop-calling, differentially enriched loops calling and loops annotation. It also contains multiple modules for interaction resolution estimation, data similarity estimation, features quantification, feature aggregation analysis, and visualization. cLoops2 with documentation and example data are open source and freely available at GitHub: https://github.com/KejiZhaoLab/cLoops2.

INTRODUCTION

Chromatin is well organized in multi-scale 3D structures as loops, domains, and compartments (1,2). These structures play critical regulatory roles for gene expression (3,4) and biological processes such as development (5) and cell cycle (6,7). Transcription factors (TFs), CTCF and cohesin, are important players in establishing chromatin architectures through the loop extrusion model (8–10). Other TFs, such as YY1 (11), Wapl (12) and ZNF143 (13), are implicated in 3D dynamic changes of the chromatin. Due to the crucial roles of chromatin interaction in cells, many efforts have been put forward to developing versatile experimental tools for elucidating the 3D structure of chromatin. Among the popular proximity-ligation-based methods, Hi-C unbiasedly detects genome-wide interactions (14,15). In contrast, ChIA-PET (16), HiChIP (17) and PLAC-seq (18) detect chromatin interactions over selected genomic regions.

Unlike other signal enrichment techniques for detecting chromatin interactions, TrAC-looping is independent of proximity ligation and detects both chromatin accessibility and chromatin interactions at high resolution (19). Therefore, it is useful for analyzing promoter-enhancer interactions. Effective computational pipelines are necessary to intercept the data to draw biological conclusions. TrAC-looping uses the DNA transposase Tn5 to capture chromatin interactions by inserting an oligonucleotide bridge between two interacting chromatin loci and thus covalently joint the two regions together for direct PCR amplification. Hi-TrAC is a streamlined version of the TrAC-looping technique and the chromatin interaction data detected by Hi-TrAC and Trac-looping have very similar features (Liu *et al.* unpublished data, data deposited to GEO with the accession number of GSE180175). Thus, in addition to provide chromatin interaction information, they also measure chromatin accessibility like ATAC-seq (20). Hi-TrAC/TrAC-looping data contain interactions between accessibility sites such as enhancers and promoters as loops and generic chromatin interactions analogous to Hi-C interactions like domains. Therefore, a comprehensive pipeline to analyze this kind of data should include: (1) preprocessing raw reads into interaction data; (2) peak-centric analysis; (3) loop-centric analysis and (4) domain-centric analysis. The peak-centric analysis methods have been well developed for ChIP-seq data (21–24), and multiple domain-centric analysis methods are available for Hi-C data (25–27). Loop-centric analysis can reveal regulatory details of *cis*-regulatory elements and loop function; however, this category of analysis methods is still underdeveloped. Ideally, a loop-centric analysis tool should consist of core modules including (1) accurate calling algorithms, (2) visualization, (3) differentially enriched loops calling analogous to the analysis of differentially expressed genes or sample-wise comparison methods, (4) loop annotations and (5) integration analysis with other data.

In this study, we introduce a new analysis pipeline, cLoops2, to address the practical analysis requirements,

*To whom correspondence should be addressed. Tel: +1 301 496 2098; Fax: +1 301 402 0971; Email: zhaok@nhlbi.nih.gov
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

especially for loop-centric analysis with preferential design for Hi-TrAC/TrAC-looping data. Based on an improved and unsupervised clustering algorithm blockDBSCAN, which is highly sensitive and unbiased, cLoops2 directly analyzes the paired-end tags (PETs) to find candidate peaks and loops. It estimates the statistical significance for the peak/loop features with a permuted local background, eliminating the bias introduced from third part peak-calling parameters tuning for calling loops. In addition to the core modules: peak-calling, loop-calling, and differentially enriched loops calling, the cLoops2 package also contains other functional modules including visualization, feature quantification, feature aggregation analysis, and loop target annotations. cLoops2 also implements many utility functions. For example, estimation of reasonable resolution for interaction contact matrix, which usually is determined semi-empirically. Although cLoops2 was designed to analyze TrAC-looping and Hi-TrAC data, it can also be applied to calling peaks from the state-of-art ChIP-seq like sequencing methods and comprehensive loop-centric analysis from ChIA-PET and HiChIP data, showing the versatility of cLoops2.

MATERIALS AND METHODS

ChIC-seq experiments

Small cell number ChIC-seq assays were performed similarly to our previous study (28). Briefly, GM12878 cells were fixed for 10 minutes with 1% formaldehyde in the DMEM medium. The reaction was stopped by adding 0.125 M glycine. Following permeabilization of 200 000 fixed GM12878 cells with 600 μl RIPA buffer at room temperature for 5 minutes, the cells were rinsed with 600 μl binding buffer twice (10 mM Tris–Cl, 1 mM EDTA, 150 mM NaCl, 0.1% TX-100) and re-suspended in 250 μl binding buffer. Before mixing with the cells, antibodies ((CTCF (Millipore, 47-729), H3K4me3 (Millipore, CS200580), H3K27ac (Abcam, ab4729), IgG (Millipore, NI03)) and PA-MNase were pre-incubated with a molecular ratio of 1:2 at 4°C for 30 min in 50 μl binding buffer for generating antibody + PA-MNase complexes. Then, the permeabilizated GM12878 cells in 50 μl binding buffer were mixed with each antibody + PA-MNase complex and incubated for 1 h at 4°C. Cells were washed using 600 μl RIPA buffer (10 mM Tris–Cl, 1 mM EDTA, 150 mM NaCl, 0.1% SDS, 0.1% NaDOC, 1% Triton X-100) four times and pelleted by centrifugation at 900g for 1 min. Next, cells were rinsed using 200 μl rinsing buffer (10 mM Tris–Cl and 10 mM NaCl, 0.1% TX-100). MNase digestion was activated by re-suspending cells in 40 μl reaction buffer (10 mM Tris–Cl, 10 mM NaCl, 0.1% TX-100, 2 mM $CaCl_2$) and incubated at 37°C for 3 min. Reactions were stopped by adding 80 μl stop buffer (20 mM Tris–Cl, 10 mM EGTA, 20 mM NaCl, 0.2% SDS). Reverse crosslinking was performed by adding 1 μl 20 mg/ml proteinase K and incubating samples at 65°C overnight. DNA was purified using the MinElute reaction cleanup kit following the manufacturer's protocol. The purified DNA was end-repaired using an End-It DNA-Repair kit (Epicenter, Cat#ER81050) and added an A base at the 3′ end using the Klenow fragment (3′→5′ exo-) (NEB, Cat#M0212L). The DNA was then ligated with Y-shaped adaptors and amplified for 16 cycles using indexing primers with the PCR condition: 98°C,10'; 67°C, 30′; 72°C, 30′ as previously reported (29,30). PCR products between 150 and 350 bp were isolated for sequencing on Illumina NovaSeq.

Peak calling algorithm of cLoops2

Candidate peak regions were obtained from blockDBSCAN with PETs distance <1 kb. If multiple parameters of *eps* and *minPts* were assigned, then blockDBSCAN with combinations of *eps* and *minPts* were performed, and candidate peaks were collected. For each candidate peak, the Poisson test was used to determine the reads' statistical significance in the peak compared to nearby regions and the same region in the control sample (input or IgG).

$$p = 1 - \sum_{i=1}^{n-1} Poisson\left(i, \max(\lambda_{ext5}, \lambda_{ext10}, \mathrm{sf} \times \max(\lambda_{bg}, pseudo), pseudo)\right)$$

In the above formula, n is the observed PETs number in the candidate peak region from the test sample. λ_{ext5} and λ_{ext10} are the mean values of observed PETs number for the upstream and downstream same size 5-folds and 10-fold windows of the candidate peak region from the test sample. In sensitive mode, they are the median values. sf is the scaling factor between the control sample and the test sample. If no control sample assigned, $\mathrm{sf} = 0$; otherwise sf is either library sequence depth ratio or the coefficient from linearly fitting of all candidate peak regions PETs number between test sample and control sample without intercept. By default coefficient from linearly fitting is used as a scaling factor and in sensitive mode, the smaller scaling factor is used. λ_{bg} is the observed PETs number in the candidate peak region from the control sample. *pseudo* is used to control empirical background noise as an adjustable parameter, by default is 1.

The candidate peak with the highest RPKM value was reported for overlapped candidate peaks with significant Poisson *P*-values if multiple *eps* and *minPts* were assigned. Finally, all *P*-values were corrected by Bonferroni correction. By default, corrected $P < 0.01$ was used to determine significant peaks. The algorithm is summarized as the cLoops2 callPeaks module.

ChIC-seq analysis

Raw paired-end reads in FASTQ format were mapped to human reference genome hg38 by Bowtie2 (31). Mapped unique paired-end reads with MAPQ ≥ 10 were used for the following analysis. Peaks called by cLoops2 were performed with key parameters of cLoops2 callPeaks -eps 100,200 -minPts 5 -bdg IgG. Default parameters were used for peak calling by MACS2 (v2.2.6) except for H3K27ac ChIC-seq, for which -board option was used. Peak calling by the SICER algorithm was performed by a faster implementation epic2 (v0.0.41) (32) with default parameters. Peak calling by HOMER was performed by findPeaks command: for CTCF ChIC-seq -style factor option was used and for H3K4me3 and H3K27ac ChIC-seq -style histone were assigned. Peak calling by SEACR was performed by both stringent mode and relaxed mode.

Comparison of peaks called by various algorithms with ENCODE peaks

To use ENCODE peaks as a reference dataset for comparing peak-calling algorithms, ENCODE peaks were quantified first in ChIC-seq data (CUT&RUN or CUT&TAG for that part of the analyses). Only the ENCODE peaks that had signal densities of 2-fold higher than flanking upstream and downstream same-sized regions and 2-fold higher than the signals in the same region in the IgG (or input) control samples were kept for the analysis. Let N_i be the number of called peaks for a tool i for one factor, n_i be the number of called peaks overlapped with ENCODE reference peaks, M be the number of ENCODE reference peaks, and m_i be the number of ENCODE reference peaks overlapped with called peaks. Then $Precision_i = n_i/N_i$, $Sensitivity_i = m_i/M_i$ and $F1_i = 2 \times \frac{Precision_i \times Sensitivity_i}{Precision_i + Sensitivity_i}$. Overlapped peaks were obtained by BEDtools intersectBed command with -f option assigned 0.5 or 0 by requiring at least half of peaks were overlapped or minimum one bp.

Loop calling algorithm of cLoops2

Candidate loops were obtained from blockDBSCAN. If multiple parameters of *eps* and *minPts* were assigned, then blockDBSCAN with combinations of *eps* and *minPts* were performed. Candidate loops were collected from multiple rounds of clustering. The hypergeometric test and the Poisson test were the same as the description in cLoops (33). The binomial test for each candidate loop was updated as following,

$$p = 1 - \sum_{k=0}^{R_{i,j}} \binom{R_i \times R_j}{k} P_{i,j}^{k} (1 - P_{i,j})^{R_i \times R_j - k}$$

where $R_{i,j}$ is the number of PETs linking the candidate anchors i, j. R_i is the number of PETs located in candidate anchor i, and R_j is the number of PETs located in candidate anchor j. $P_{i,j}$ is the possibility of observing 1 PET link of the two regions, estimated from local permutation background regions as $P_{i,j} = \frac{\sum_{\hat{i},\hat{j}} \frac{R_{\hat{i},\hat{j}}}{D_{\hat{i}} \times D_{\hat{j}}}}{n_{i,j}}$. $R_{\hat{i},\hat{j}}$ is the number of PETs linking the local permutation regions. $D_{\hat{i}}$or $D_{\hat{j}}$ is the number of PETs in the permutation anchor $\hat{i}$ or $\hat{j}$. $n_{i,j}$ is the total number of permutation loops. The changes generated similar trend binomial P-values with the previous method but removed the total number of PETs from the math formula.

Aggregation analysis of loops

An 11 × 11 contact matrix was constructed for a loop from interacting PETs, together with its upstream and downstream same size 5 windows. An individual enrichment score for a loop was calculated as the number of PETs in the matrix center divided by all others' mean values. The global enrichment score was the mean value of all enrichment scores for individual loops. Heatmap was plotted of the average matrix for all normalized 11 × 11 matrices. In contrast to the aggregation analysis of loops in Juicer (34), which by default filters some close loops before aggregation, cLoops2 runs the analysis for all loops by default.

TrAC-looping data processing

Raw reads in FASTQ files were processed by tracPre.py in the cLoops2 package mapped to human genome hg38 into BEDPE files. After processing into cLoops2 data by the cLoops2 pre module, cLoops2 samplePETs were used to do sub-sampling for the two samples into 100 million PETs for a fair comparison of loop calling and differentially enriched loop calling. Peaks were called by the cLoops2 callPeaks module with key parameters of -eps 100 -minPts 20 -sen. Loops were called by cLoops2 callLoops module with parameters settings of -eps 200,500,1000 -minPts 5 -cut 1000. The cLoops2 callDiffLoops module called differentially enriched loops with default parameters settings. Raw reads of RNA-seq data were mapped to hg38 by STAR (v2.7.3a) (35), and fold changes of activated to resting CD4 + cells were obtained by Cuffdiff (v2.2.1) (36).

Public data used

Re-analyzed public datasets of raw FASTQ files, mapped reads BED files, and ENCODE peaks were summarized in Supplemental Information.

RESULTS

Overview of cLoops2

cLoops2 is a full suite of tools for 3D genomic interaction data, improved from our previous work cLoops (33), with many extended function modules (Supplemental Figure S1 and Figure 1). Like cLoops, cLoops2 is also a lightweight command-line tool for Linux and macOS for use via terminal and Python, which can be easily integrated as a backend component for potentially more sophisticated tools with web or graphical interfaces. The learning curve of cLoops2 is comparable to other tools like SAMtools (37), BEDtools (38) or MACS2 (21). The main modules of cLoops2, such as peak/loop/domain calling algorithms and features aggregation analysis, are integrated into the main command, and can be executed through the prefix cLoops2 (Supplemental Figure S1). Meanwhile, extendable supplementary analysis scripts can be executed independently. For example, tracPre2.py processes raw Hi-TrAC (tracPre.py processes TrAC-looping) data from FASTQ files to uniquely high-quality mapped paired-end tags (PETs) BEDPE files (Supplemental Figure S1). Each sub-command can be run through the 'cLoops2 sub-command' (for example, 'cLoops2 callPeaks'), with a program description, parameter details, and examples shown as the default output.

The core algorithm for peak-calling and loop-calling in cLoops2 was based on a new clustering algorithm called blockDBSCAN (Figure 1A), which further improved cDBSCAN in cLoops. Each PET from the mapping result was marked as (x, y) for the genome coordinates, with x marking the smaller coordinate and y marking the bigger one for each sequencing tag. There are peaks in the 1D linear space and loops in the 2D space with some random noise in both for ideal Hi-TrAC/TrAC-looping, ChIA-PET (16),

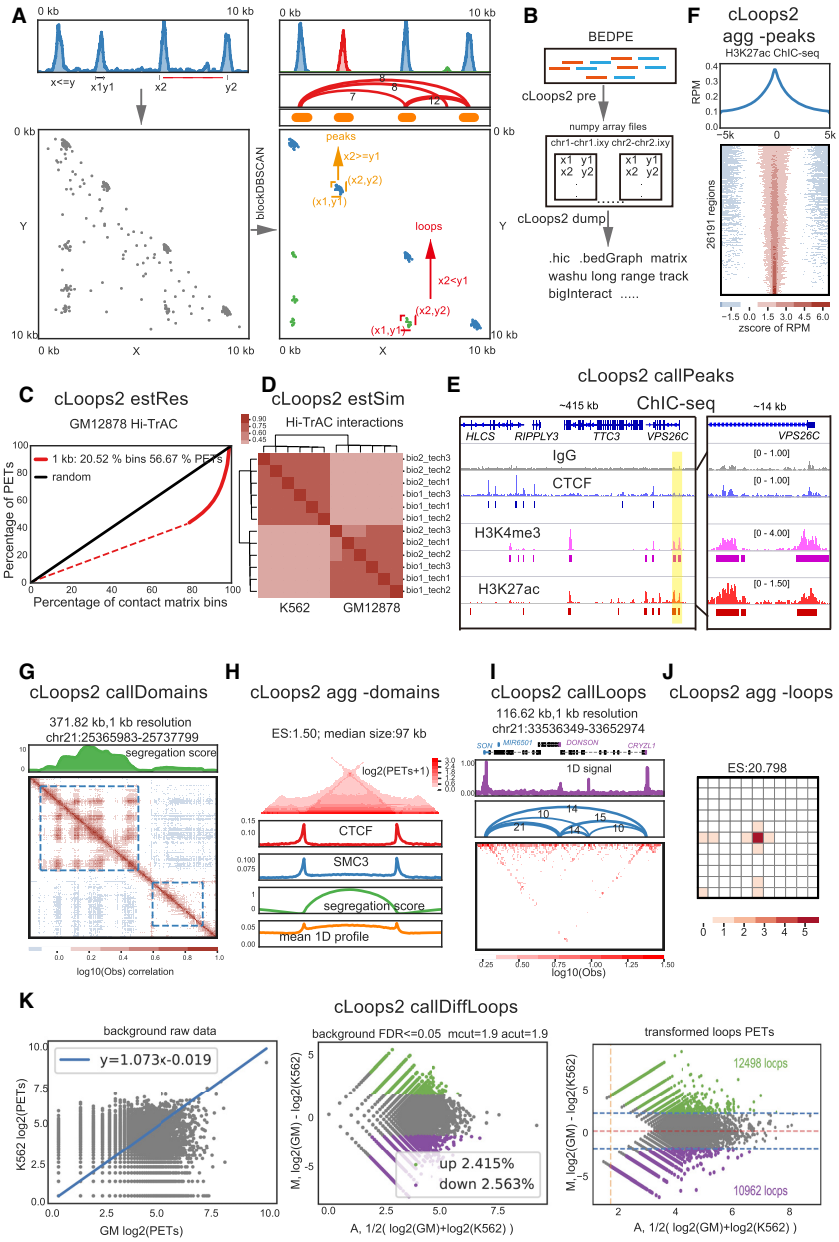

Figure 1. Overview of cLoops2 module features. (**A**) A demo of cLoops2 peak-calling and loop-calling core algorithms based on blockDBSCAN clustering with hypothetical data. There were peaks, loops, and background noise in the hypothetical data as they mimicked the properties of Hi-TrAC/TrAC-looping or HiChIP data. Processed paired-end tags (PETs) from sequencing data were marked as (x, y) for each one for their mapped coordinates in the genome. When PETs were projected into 2D space, close PETs (such as ($x1$, $y1$)) were projected to a diagonal line or nearby, while distal PETs (such as ($x2$, $y2$)) were projected away from the diagonal line. After blockDBSCAN clustering, candidate peaks and loops were identified from the clusters. (**B**) The cLoops2 pre and dump modules for data input-output stream. The cLoops2 pre module processes mapped PETs from BEDPE files into cLoops2 specific data files of Numpy array files. The cLooops2 dump module converts cLoops2 data to other data types such as HIC files. (**C**) An example of the cLoops2 estRes module for estimating reasonable contact matrix resolution for GM12878 Hi-TrAC data. The dashed line shows the bins from the contact matrix with only singleton PET, and the solid lines show the bins with multiple PETs. The random line indicates all PETs are evenly distributed in the contact matrix bins. We defined the highest interaction resolution for which there are more than 50% non-singleton PETs in the defined bin sizes (as here for 1kb bin size). (**D**) An example of the cLoops2 estSim module for estimating interaction similarities with 1kb resolution among Hi-TrAC samples. (**E**) An example of the cLoops2 callPeaks module for peak-calling of ChIC-seq data. The left panel shows a randomly selected genomic region with the ChIC-seq signals for IgG, CTCF, H3K4me3, and H3K27ac in GM12878 cells and called peaks; the right panel shows the zoom-in region for a better view of captured peaks. (**F**) An example of the cLoops2 agg module for showing aggregated peaks of GM12878 H3K27ac ChIC-seq data. (**G**) An example of cLoops2 callDomains module for domain-calling from the GM12878 Hi-TrAC data with key settings of -bs 1000 -ws 100 000. Positive segregation scores indicate domain regions. (**H**) An example of the cLoops2 agg module for showing aggregated domains of GM12878 Hi-TrAC data. (**I**) An example of the cLoops2 callLoops module for loop-calling from the GM12878 Hi-TrAC data. The cLoops2 plot module generated the plot. (**J**) An example of the cLoops2 agg module for showing aggregated loops of the GM12878 Hi-TrAC data. (**K**) Examples of output figures for cLoops2 callDiffLoops module comparing Hi-TrAC loops from GM12878 and K562 cells. Nearby loop background data were used to fit a transformation line between two datasets (left panel). MA plot of background data was used to get the cutoffs for average and fold change with FDR $\leq$ 0.05 (middle panel). The transformation line and cutoffs estimated from the background were then applied to loops to find significantly enriched loops (right panel).

and HiChIP (17) data (Figure 1A). The PETs with short distance (such as (x_1, y_1) from peaks) in the 1D space were projected to the diagonal line in the 2D space, while the long-range interaction PETs (such as (x_2, y_2) from loops) were projected to the space distant from the diagonal line. After blockDBSCAN clustering, noisy points were removed (gray dots in Figure 1A), and the number of clusters was automatically determined. Therefore, every cluster can be marked as [(x_1, x_2), (y_1, y_2)], with x_1 marking the minimal coordinate and x_2 marking the maximal coordinate of the left end tags; y_1 marking the minimal coordinate and y_2 marking the maximal coordinate of the right end tags. Based on the boundary overlaps of x_2 and y_1, clusters can be classified into two groups, candidate peaks ($x_2 \geq y_1$) and candidate loops ($x_2 < y_1$). These candidate peaks and loops were tested against the permutation local background to obtain statistical significance to further determine whether they were putative peaks or loops.

Main modules of cLoops2

The cLoops2 pre module preprocesses the input BEDPE format data of mapped PETs into cLoops2 specific data files as a folder, of which each interacting chromosome pair is saved into a file. All other cLoops2 functions start from this data folder. Each file is a Numpy (39) 2D array, recording the mapped coordinates of PETs. The cLoops2 dump module converts cLoops2 specific data into popular supported file types that can be loaded in Juicebox (40) (HIC file), the WashU Epigenome Browser (41) (long-range interaction track file), and the contact matrix in a text file that can be loaded into Python, R for further analysis, or TreeView (42) for visualization (Figure 1B). One advantage of the cLoops2 data structure is that the data can be converted back to interacting PETs (cLoops2 dump -bedpe) while other matrix-based data formats such as HIC or cooler (43) lose the PETs-level information, which limits them to show only matrix heatmap based visualization with limited resolutions but not the data with any resolutions, 1D signal or interacting PETs as arches.

The cLoops2 estRes module estimates the resolution of an interaction contact matrix based on singleton PETs and bins (Figure 1C). For a specific input resolution, the cLoops2 estRes module assigns each PET into a contact matrix bin. Accumulation of PETs (Y-axis) against the accumulation bins (X-axis, bins sorted by PETs in the bin ascendingly) for the input resolution was plotted to determine the interaction signal enrichment. If PETs are evenly distributed between genomic locations, there will be a straight diagonal line (black line and red dash line in Figure 1C), which is true if we shuffled the two ends of all PETs to an expected background. Otherwise, PETs will be enriched in only a few bins if there are only a few strong interactions. The curve will then show a prominent and steep rise towards the higher ranked bins (solid red line in Figure 1C). There are high possibilities that those singleton PET bins are due to noise or background signals. Because of the existence of chromatin domains and loops in the 3D genome, specific PETs should display certain degrees of enrichment. Therefore, we define the highest genome-wide resolution as more than 50% PETs (solid curves in Figure 1C) detected in multiple PET bins. The estimation is only a whole-genome estimation, and some local regions may have higher or lower resolutions. The cLoops2 estRes module can also estimate similarities among different experimental interaction samples (Figure 1D).

There are four main callers for features finding in cLoops2: callPeaks for interaction data such as HiTrAC/TrAC-looping data or 1D genome feature profiling data such as ChIC-seq (Figure 1E), callDomains (Figure 1G), callLoops (Figure 1I) and callDiffLoops for calling differentially enriched loops (Figure 1K) from HiTrAC/TrAC-looping data. For the identified features, versatile aggregation analysis were also implemented in the cLoops2 agg module for peaks (Figure 1F), domains (Figure 1H), and loops (Figure 1J). Meanwhile, the cLoops2 quant module can generate similar files to the output of callers if features are called in one sample and quantified in another sample (Supplemental Figure S1).

To call differentially enriched loops between two conditions, loops from pairwise samples are combined first and then quantified in both samples. Loops' nearby regions are defined as background, quantified, and then fitted linearly (Figure 1K, left panel). False discovery rate (FDR) is a required parameter, which is set to 0.05 by default, to find the cutoffs of average and fold change in an MA plot for background data (Figure 1K, middle panel). The fitted slope and intercept transform the PETs in loops of treatment dataset to control dataset, assuming there should be no difference in background data (Figure 1K right panel). The average and fold change cutoffs drawn from the background data are then applied to transformed loops data (Figure 1K, right panel). Poisson P-value is assigned to each loop as $p = 1 - \sum_{i=1}^{fg-1} Poisson(i, \max(bg, fgNearby, bgNearby, pseudo))$, where fg stands for the bigger value of PETs in the testing loop; bg stands for the smaller value; fgNearby is the number of PETs from background data for the testing condition, and bgNearby is the number of PETs for background data from the other condition. Pseudo is a general noise control value set to 1 as a default and can be adjusted through its parameter. The Poisson P-values are further corrected by the Bonferroni correction.

The blockDBSCAN algorithm

In a previous study of cLoops, we proposed cDBSCAN, which was improved from the popular DBSCAN algorithm, for clustering interacting PETs directly to find candidate loops (33). Analyzing PETs directly, but not the contact matrix bins, achieves resolution scale-free performance, is an important feature to identify loops' variable sizes and accurate boundaries. Although cDBSCAN worked well for previously tested ChIA-PET, TrAC-looping, HiChIP and Hi-C data, it is quite time-consuming for deeply sequenced libraries (44). Therefore, we further improved the algorithm's speed and named it blockDBSCAN (Figure 2). There are two key parameters in blockDBSCAN with the same meaning as those in DBSCAN, eps and minPts : eps defines the maximum distance that two points are neighbors and minPts defines a cluster's minimal number of points. Unless specifically mentioned, the distance measurement used in blockDBSCAN refers to Manhattan distance.

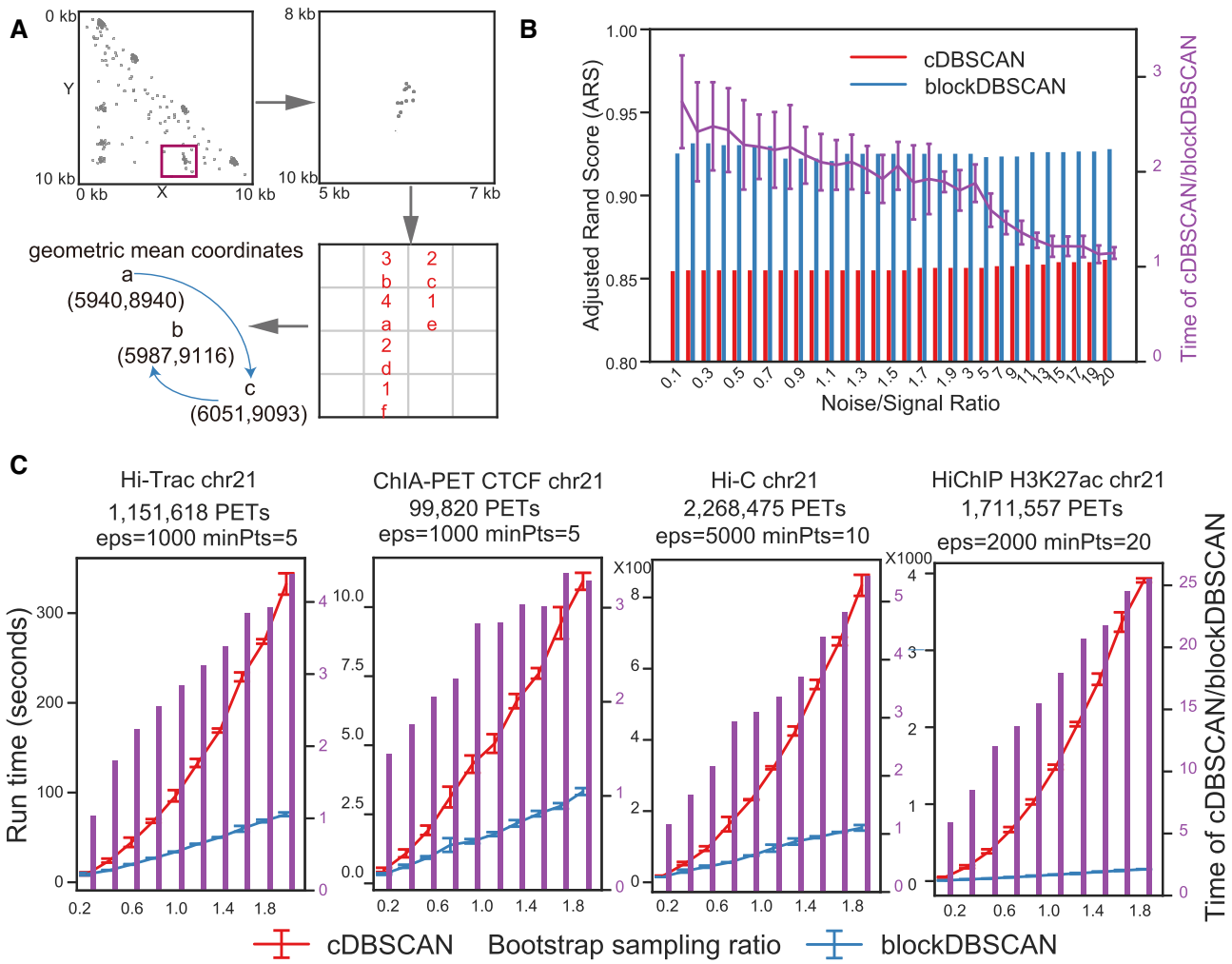

Figure 2. The blockDBSCAN algorithm. (**A**) A demo of blockDBSCAN algorithm processes on hypothetical data. The red rectangle highlights the region for a detailed demonstration. After indexing and noise removal, informative PETs were located in the indexed blocks with the side length of eps (one key parameter of DBSCAN). For each indexed block, marked as a to f, PET numbers in them were also annotated. Geometric mean coordinates were computed for each indexed block from PETs in that indexed block for the following clustering at the block level. (**B**) The comparison of running CPU time and accuracy of cDBSCAN and blockDBSCAN at different noise/signal ratios was based on five repeats of simulation data. The left *y*-axis marked the bars for adjusted rand scores (ARS), showing consistency with simulation data true labels; the right y-axis marked the error bars for running time ratios. ARS measured the similarity between clustering results and true labels ranging from –1.0 to 1.0, with 0 indicating random labeling and 1 is a perfect match. The simulation data were generated in the same way described in cLoops (33). (**C**) Comparison of running CPU time of cDBSCAN and blockDBSCAN with real interaction data of chromosome 21 from Hi-TrAC, CTCF ChIA-PET, *in situ* Hi-C and H3K27ac HiChIP data in GM12878 cells (Supplemental Information), based on 5 repeats and sampling.

Same to cDBSCAN, all PETs are projected to index blocks with the size of eps (Figure 2A). In cDBSCAN, after indexing and noise removal, clustering is performed comparable to DBSCAN for all remaining PETs but with a limited search to neighbor index blocks. blockDBSCAN improves the last step by marking each index block as a point with weight, with the coordinate being the geometric mean of PETs, and the weight being the number of PETs in that index block (marked from a to f in Figure 2A). When expanding the cluster from one index a to another index b, the distance between the two index blocks' coordinates will be calculated first. If the distance is less than eps and together more PETs than minPts, index b is clustered. Otherwise, if the distance is larger than eps, then distances of each pair of points between index a and b will be calculated to find if there is any pair distance less than eps. If there are more PETs than minPts, index b is also clustered with index a. In any other condition, index *b* will not be clustered together from index a. Python implementation of the algorithm close to pseudo-code can be found at https://github.com/YaqiangCao/cLoops2/blob/master/cLoops2/blockDBSCAN.py. Clustering performed at the level of index blocks can effectively reduce algorithm complexity, and therefore increase clustering speed. We evaluated the performance of blockDBSCAN against cDBSCAN with simulated data, and confirmed that blockDBSCAN was faster than cDBSCAN, and even a little better matched with true simulated labels of signals measured by the adjusting rand index (ARS) (Figure 2B). We further validated the clustering algorithm efficacy for finding candi-

date loops with real Hi-TrAC data, CTCF ChIA-PET data, Hi-C data, and H3K27ac HiChIP data (Figure 2C), achieving about 2–15-fold speed improvement and with nearly the same memory usage (Supplemental Figure S2). blockDBSCAN was estimated approximate $O(N)$ complexity as run time increases nearly linearly with smaller slopes comparing to cDBSCAN when the number of PETs increases (Figure 2C).

Peak-calling for next-generation ChIP-seq like data

Peak-calling is a fundamental analysis step for popular 1D genomic feature profiling methods such as ChIP-seq (45) and ATAC-seq (20). Currently, the most popular peak-calling algorithms were designed for ChIP-seq, such as MACS (21). However, classical tools may have limitations such as reduced precision (46) for the latest emerging technologies such as CUT&RUN (47) and ChIC-seq (48). A possible explanation for that is a different signal-to-noise ratio and variable peak sizes compared to ChIP-seq. Loop-calling tools, such as mango (49) for ChIA-PET data and hichipper (50) for HiChIP data, usually start from peak-calling, and then combinations of peaks are used to find candidate loops. It is inferable that inaccurate or insensitive peak-calling may lead to false positive or true negative candidate loops for those tools. Meanwhile, it was also noticed the peak-calling results from both FitHiChIP (51) and hichipper (50) were strongly biased due to HiChIP specific biases (52), which may bias the accuracy of finally called loops. Thus, to date, generalist peak-calling algorithms are still needed. In a previous study of cLoops (33), we only used cDBSCAN to identify candidate loops. Now we found that the density approaching principle can also be applied to identifying candidate peaks (Figure 1A).

To benchmark the performance of the cLoops2 callPeaks module, we generated ChIC-seq datasets for the transcription factor CTCF with sharp peaks (Figure 3A), the histone modification H3K4me3 with sharp peaks (Supplemental Figure S3A), the histone modification H3K27ac with both sharp and broad peaks (Figure 3B), and IgG ChIC-seq control data. The peaks called from these datasets were compared with three popular ChIP-seq peak callers: MACS2 (21), SICER (53), HOMER (54) and SEACR (46), which was designed for CUT&RUN protocol (Materials and Methods).

We first examined the examples of called peaks from the CTCF ChIC-seq data and compared them to the ENCODE reference peaks from ChIP-seq (Supplemental Information, Materials and Methods). We noticed that cLoops2 captured most of the significant peaks in three randomly chosen regions, which were well-aligned with the ENCODE reference peaks (Figure 3A). The density-based approach also found accurate boundaries of peaks (Figure 3A). The peaks called by both SEACR's stringent and relaxed modes were too broad. The SEACR stringent mode also missed some significant peaks, resulting in a limited total number of peaks found (Figure 3A). The peaks called by SICER were too broad for transcription factor peaks such as CTCF binding sites using default parameters without optimization tunning (Figure 3A), although it has the potential to work well with more specific parameter settings for ChIC-seq data (55). MACS2 yielded the most peaks in replicate 2 (44,714); however, many of them may be background regions located near true peaks (Figure 3A). Meanwhile, for MACS2, fewer peaks were called for replicate 1 (31 880); the difference in numbers of peaks in the two replicates resulted in lower consistency between replicates (Figure 3A and Supplemental Figure S3C). HOMER achieved similar performance as cLoops2 for the CTCF data (Figure 3A). Compared to the cLoops2 results, peaks called by HOMER were sharper, which were apparently much narrower than the ChIC-seq peaks as shown by the Genome Browse images, indicating improper capturing of peak boundaries.

Regarding the peak calling from the H3K4me3 ChIC-seq data, cLoops2 also worked well (Supplemental Figure S3A). Both the SEACR stringent and relaxed modes missed some significant peaks, resulting in much lower numbers of peaks as compared to the ENCODE peaks (Supplemental Figure S3A). The default SICER setting resulted in some peaks stitched together (Supplemental Figure S3A). MACS2 called the most peaks for both replicates; again, many of them may be just background signals near true peaks. HOMER achieved similar performance with cLoops2 for calling H3K4me3 peaks (Supplemental Figure S3A).

H3K27ac displays both sharp and broad peaks in the genome (Figure 3B). cLoops2 detected most of the significant reference peaks, called similar peaks as reference peaks, and showed high consistency between two biological replicates (Figure 3B). Both SEACR stringent and relaxed modes called too many peaks and were too sensitive for replicate 1 (Figure 3B). Similar to SEACR, SICER was too sensitive and called too many broad peaks (Figure 3B). MACS2 and HOMER called peaks from many background regions near highly enriched peaks (Figure 3B).

Precision, sensitivity, and the F1 score (56) were used as metrics to evaluate the performance of the genome-wide peak-calling methods by comparing the discovered peaks with the ENCODE reference peaks (Materials and Methods). If only requiring a minimal overlap of 1 bp between called peaks and reference peaks, then broad peaks overestimated of precision and sensitivity (Supplemental Figure S3B) in that only small parts of the called peaks overlapped with reference peaks at a low overlapping ratio. Therefore, we required half of the candidate peaks to overlap with the reference peaks to get evaluation metrics. With this requirement, cLoops2 had the highest F1 scores for all the three ChIC-seq datasets, showing both high precision and sensitivity for calling accurate peaks (Figure 3C). cLoops2 also showed comparable reproducibility between replicates (Supplemental Figure S3C) and correlation between gene expression and the peak density at the transcription start sites (Supplemental Figure S3D). Assuming signals from the same region of IgG control of a putative peak and the putative peak's flanking regions are background. We can use the signal-to-noise ratio distribution to check whether putative peaks are properly called. Misclassification of background regions as putative peaks or vice versa will both decrease the distribution of signal-to-noise ratios. Further comparison of signal densities in peaks against their flanking upstream and downstream regions of the same size (Figure 3D) and against the same regions in

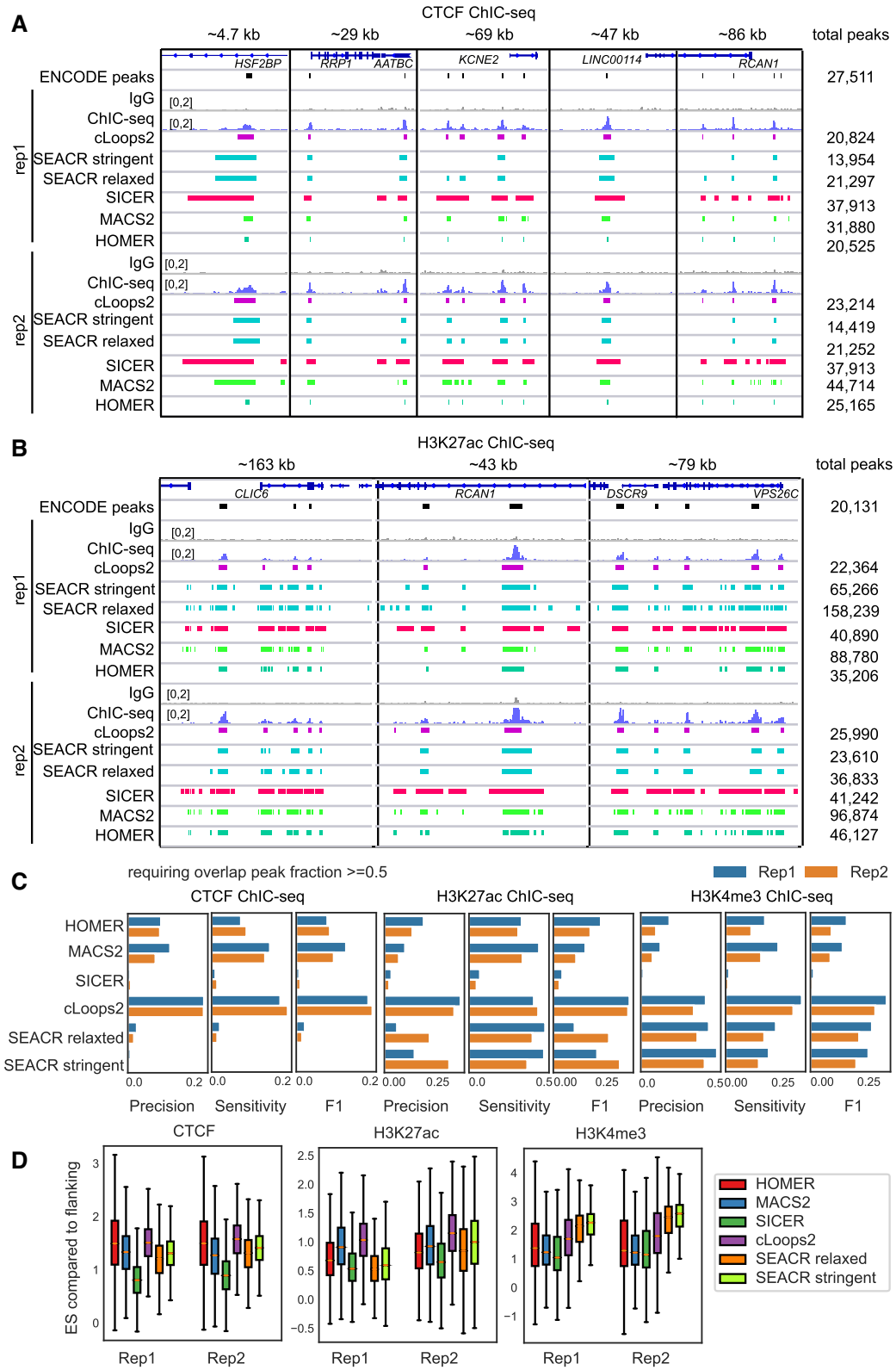

Figure 3. Peak calling by cLoops2 from ChIC-seq data. (**A**) Genome Browser images showing the CTCF ChIC-seq data and peaks called by different tools. Two biological replicates of CTCF and IgG ChIC-seq datasets were shown. (**B**) Genome Browser images showing the H3K27ac ChIC-seq data and peaks called by different tools. (**C**) Precision, sensitivity, and *F*1 scores for peaks were called by different tools by comparing with the ENCODE peaks (Materials and Methods), requiring peaks overlap fractions ≥0.5. (**D**) ChIC-seq signal-to-noise enrichment score (ES) from various tools. The ES was calculated by comparing signals in the peak region with that in the flanking upstream and downstream same-sized regions.

the IgG control sample (Supplemental Figure S3E), especially for CTCF and H3K27ac, indicates that cLoops2 performed the best to capture peaks with proper boundaries.

cLoops2 also demonstrated excellent performance for peak calling from CTCF and H3K27ac CUT&RUN data, as shown by the Genome Browser snapshots of randomly selected examples (Supplemental Figure S4A and B) and the whole genome evaluation metrics (Supplemental Figure S5). Due to the differences between the ENCODE reference peaks of H3K27me3 and CUT&RUN data, it is currently a challenge to carry out a fair comparison of peak-calling performance for H3K27me3 (Supplemental Figure S4C). cLoops2 also achieved the best performance for H2A.Z, H3K4me1 and H3K27ac CUT&TAG data measured by F1 scores compared to the corresponding ENCODE reference peaks (Supplemental Figure S6). It should be noted that SEACR was designed for CUT&RUN series technology with low background, while cLoops2 was designed as a general peak-caller based on the density approaching principle. However, even for the CUT&RUN series data, cLoops2 works well and, to some extent, better than SEACR, indicating that the principles of the cLoops2 peak-calling algorithm can be applied to potentially more genomic features profiling technologies under development.

Comprehensive analysis of loops from chromatin interaction data

Many tools have been proposed for calling loops for Hi-C (such as HiCCUPS (34) and SIP (44)) and HiChIP (such as hichipper (50) and FitHiChIP (51)). Most of them were specifically designed for only one limited data type, restricted to resolution settings, and lacking downstream analysis modules. Meanwhile, cLoops can call accurate loops for a broad range of interaction data (such as ChIA-PET, Hi-C, HiChIP and TrAC-looping) without the limitation of specific resolutions (33). Therefore, we extended the core of the loop-calling algorithm in cLoops to cLoops2 with more loop-centric analysis.

TrAC-looping data (19) were used to show further the performance of cLoops2 in loop-calling and identifying differential loops between different samples. Even without the explicit step of peak-calling for loop-calling, >70% of loops have at least one anchor overlapping with peaks (Supplemental Figure S7A and B). Those loops without any anchor overlapped with peaks also show highly enriched interaction signals comparing to nearby regions (Supplemental Figure S7C), suggesting that some putative loop anchors do not need to be strong peaks. This was exemplified by a strong loop (Supplemental Figure S7D), one of whose anchors was not a peak. We also tested cLoops2's performance on peak-calling and loop-calling from the resting CD4+ T cells TrAC-looping data by sub-sampling to check the dependence on the library sequencing depth (Supplemental Figure S8). Reproducibilities among sub-sampling replicates of peak-calling are better than loop-calling and become steady when there are >30 million final PETs, achieving >85% overlapped peaks (Supplemental Figure S8). Precisions for both peak-calling and loop-calling are high (≥90%) for detecting reproducible peaks or loops comparing to all PETs for sub-sampling datasets. Meanwhile, sensitivities are dependent on PETs number, and therefore F1-scores are dependent on the PETs number (Supplemental Figure S8).

As exemplified with a 17.8kb genomic region, which contained the *IRF2BP2* gene, cLoops2 detected complex and different looping structures between resting and activated CD4+ T cells (Supplemental Figure S9A). Markedly more loops were identified around the *IL2* gene, which is a critical gene required for T cell proliferation (57), in the activated CD4+ T cells than in the resting CD4+ T cells (Supplemental Figure S9B), indicating that cLoops2 can identify dynamic loops associated with cellular functions. Aggregation analysis of cell-specific loops and their nearby regions revealed 11 097 highly enriched loops in resting CD4+ cells and 14,229 in activated cells, respectively (Figure 4B). Both sets of the specific loops exhibited a lower or similar level of signals than the nearby regions in the other cell type (Figure 4B). For these cell-specific loops, we further examined their anchors of TrAC-looping 1D signals, and found only a few of the anchors (1,578) were shared by the two cell types while most of them were cell-specific (15 068 in resting CD4+ cells and 23 ii148 in activated CD4+ cells) (Figure 4C). The observation that most anchors of the cell-specific loops were also cell-specific suggests two interesting possibilities: (i) cell-specific accessibility was guided by cell-specific looping structures and (2) cell-specific looping was guided by cell-specific TFs, which can be detected by 1D profiling methods such as ChIP-seq, DNase-seq and ATAC-seq.

We further annotated these cell-specific loops by the cLoops2 anaLoops module to identify their target genes and examine the target gene expression (Figure 4D). If a gene promoter looped with multiple enhancers and was associated with distinct loops in different cell types, it was termed an alternative target. Otherwise, if a gene promoter had loops only in one cell type, it was defined as either resting or activated CD4+ T cell-specific target gene. Globally, resting or activated CD4+ T cell-specific loops targeted specific genes had a slightly higher expression, but not significantly, in the corresponding condition (Figure 4D). For the alternative target genes, even though they did not show significant global expression differences between the two cell types, many of them used different enhancers between the resting and activated CD4+ T cells, which were enriched in lymphocyte differentiation and activation pathways (Figure 4E). The *IRF8* gene is a critical regulator of immune function. Its expression was not detectible and no significant loops originating from the *IRF8* promoter were identified in the resting CD4+ T cells (Figure 4F). In the activated CD4+ T cells, the *IRF8* gene locus exhibited complex looping structures, including both promoter–enhancer and enhancer-enhancer loops, accompanied by high levels of *IRF8* expression (Figure 4F). E1 and E2 were looped together; E2 was looped with the promoters of both the *IRF8* and *AC092723.5* genes. This example showed clear evidence of gene expression regulation by enhancer-promoter loops upon T cell activation. On the other hand, the regulation of alternative looping target genes might be much more complex and need further study. For example, the *ANXA1* gene was expressed at a similar level in the resting and activated CD4 + T cells (Figure 4G). Multiple

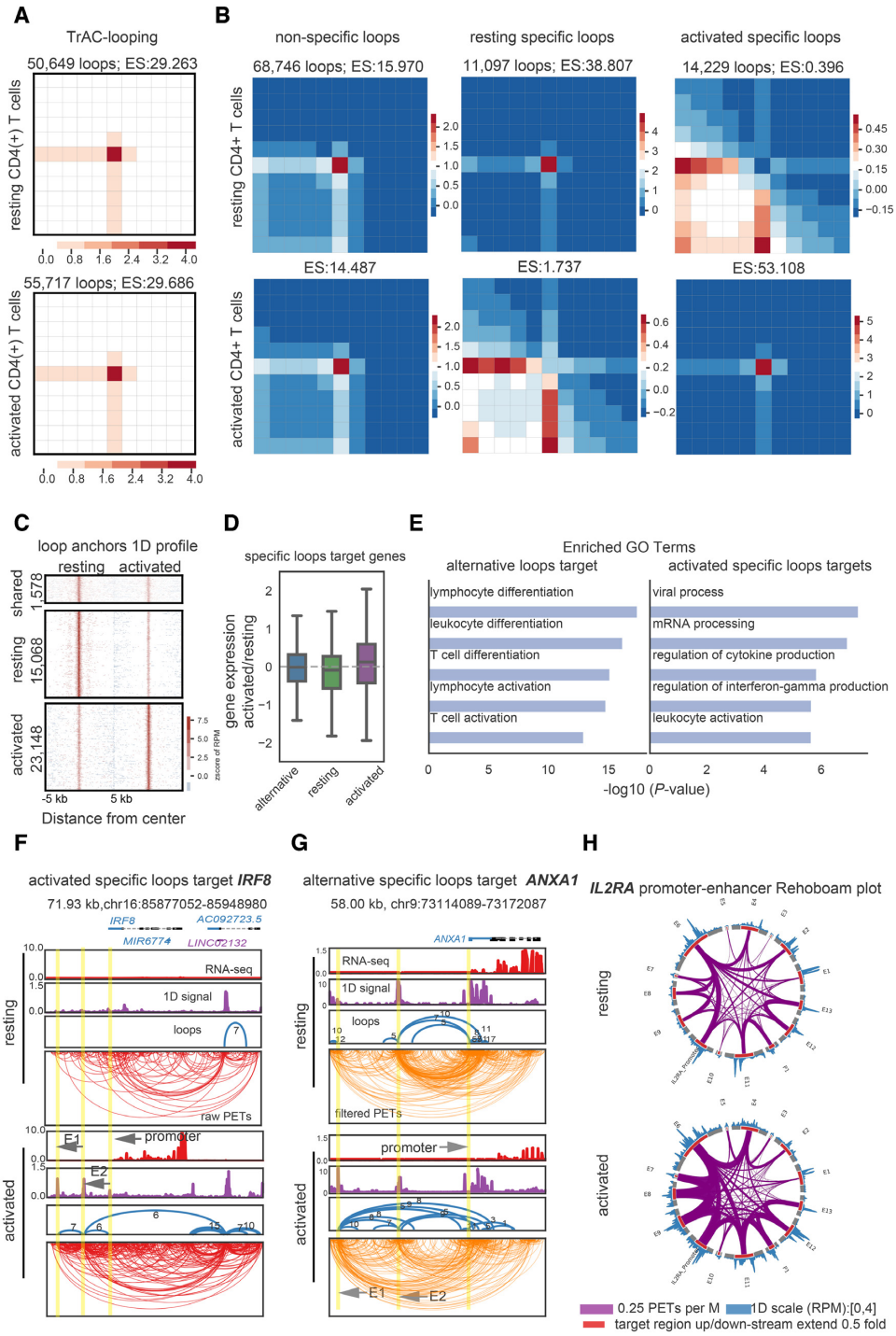

Figure 4. cLoops2 calls differentially enriched loops from TrAC-looping data. (**A**) Aggregation analysis of called loops from TrAC-looping data of resting and activated CD4+ T cells. Mapped cis-PETs were sub-sampled to 100 million for analysis and comparison. Enrichment score (ES) was the average value of the matrix center divided by others of all loops. (**B**) Aggregation analysis of non-specific and specific loops between resting and activated CD4+ T cells. The cLoops2 callDiffLoops carried out the analysis. (**C**) Aggregation analysis of TrAC-looping 1D signals on combined anchors from cell-specific loops. Only limited anchors were shared between different cells. The specific 1D signals also indicated cell-specific accessibilities of the anchors. (**D**) Gene expression associations with cell-specific loops. Alternative means a gene's promoter is looping by different enhancers in two conditions. (**E**) Enriched gene ontology (GO) terms for cell-specific loops targeting genes. Only the top 5 were shown. (**F**) Example of activated T cell-specific loops targeting gene *IRF8*. Blue color marks the first exon of the gene in the positive strand, and purple color marks the first exon of the gene in the negative strand. E1 and E2 mark the enhancers that were active only in the activated CD4 T cells. (**G**) Example of alternative specific loops targeting gene *ANXA1*. Only PETs overlapped with any end of loop anchors were remained, performed with the cLoops2 filterPETs module. (**H**) Example of Rehoboam plots of interacting enhancers and promoters of the *IL2RA* gene locus. Putative enhancers were marked as E1 to E13, inferred from loop anchors interacting with *IL2RA*'s promoter from the two conditions (Supplemental Figure S10). A Rehoboam plot can be obtained through the cLoops2 montage module. TrAC-looping 1D profiles were shown outside the circle, and the levels of interaction were shown as widths of arches that link the interacting regions.

enhancers were looped with the *ANXA1* gene promoter in the resting and activated CD4 T cells. While E1 had stronger interaction signals with the *ANXA1* promoter in the activated CD4+ T cells, E2 showed stronger interactions in the resting CD4+ T cells. This example demonstrated that even for the non-differentially expressed genes in different cell types, the regulatory mechanisms by enhancers might be different.

The quantitative measurement of interactions among transcriptional regulatory elements such as enhancers and promoters is important to understand the mechanisms of gene expression regulation. However, the current interaction heatmaps neither provide such quantitative information nor reveal the fine looping structures, especially when comparing different cell types or conditions. For example, the *IL2RA* gene promoter interacted with a total of 13 potential enhancer regions in the resting and activated CD4+ T cells together (Supplemental Figure S10). It was difficult to see quantitative changes from the interaction heatmaps (Supplemental Figure S10). Although the arch loop plots showed overall changes of interaction patterns at this genomic locus, it was difficult to directly visualize the dynamic changes of promoter interactions (Supplemental Figure S10). To facilitate direct visualization of changes in promoter-enhancer interactions, we introduced the Rehoboam plots (Figure 4H). In a Rehoboam plot, each target region with extended nearby regions was shown as a part of a circle, TrAC-looping 1D profiles were shown outside the circle, and interaction signal levels were shown as the widths of arches between interacting regions. With the Rehoboam plots, it was easy to conclude that (1) globally more interactions were observed with the *IL2RA* promoter in activated CD4+ T cells; (2) E6 and E11 were looped together in activated CD4+ T cells but not in resting CD4+ T cells; and (3) E9 and E12 had much higher accessibility in activated CD4+ T cells (Figure 4H). Thus, a Rehoboam plot could serve as a useful tool to visualize dynamic changes of interactions and accessibilities of regulatory regions for a specific locus of interest.

We also demonstrated the performance of loop calling and differentially enriched loop calling by cLoops2 from the latest RAD21 ChIA-PET data (58) (Supplemental Figure S11) and H3K27ac HiChIP data (59) (Supplemental Figures S12–S14). The anchors from both the RAD21 specific loops and H3K27ac specific loops showed high cell-type specificity (Supplemental Figures S11C and S14C). Higher gene expression levels were also observed for those target-specific loops (Supplemental Figures S11D and S14D). As RAD21 or H3K27ac guides these loops, we also observed a high binding difference from the ChIP-seq data at loop anchors (Supplemental Figures S11E and S14E). These results indicate that the application of cLoops2 could be extended to different datasets such as ChIA-PET and HiChIP.

Additionally, cLoops2 called comparable loops with state-of-art methods designed explicitly for HiChIP data such as FitHiChIP (51), hichipper (50), mango (49), and MAPS (60) from the GM12878 H3K27ac HiChIP data (Supplemental Figures S12–S13 and Supplemental Information). Among these tools, only cLoops2 implements more downstream analysis modules, while others are only mostly limited to loop-calling. However, we think it is still a challenge to call accurate loops from H3K27ac HiChIP data. Currently, no tool outperforms others from all aspects.

GWAS SNPs target annotation from loops

It has been recognized for a long time that most of the disease or trait-associated single-nucleotide polymorphisms (SNPs) identified from genome-wide association studies (GWAS) are in non-coding regions in the human genome (61,62). Linking the associations to functional causality requires accurately annotating the target genes for the SNPs in the current post-GWAS era (63,64). Since cLoops2 is capable of annotating enhancers to their target genes from TrAC-looping or Hi-TrAC data, it can facilitate the annotation of target genes for the SNPs located within loop anchors. For a total of 116,411 SNPs annotated in the NHGRI-EBI GWAS Catalog (62), 11 995 (10.3%) of them were overlapped with the loop anchors identified in GM12878 cells, which is significantly higher than the expected background (Figure 5A). More than 6000 of the overlapped SNPs were located in the enhancer regions of loop anchors (Figure 5B). Overall, for the SNPs mapped to the loop anchors, a SNP had an average of 1.68 direct target genes, and had 3.37 indirect target genes in the loop network by average (Figure 5C). In NHGRI-EBI GWAS Catalog annotations, we noticed that if a SNP was located in long non-coding (lncRNA), its target gene was annotated as lncRNA. However, we found that many lncRNAs were located in loop anchors and thus they may regulate target gene expression by chromatin looping. We showed here three examples of lncRNA genes hosting SNPs, which were looped to promoters or enhancers detected by Hi-TrAC (Figure 5D–F) but not shown from GM12878 RAD21 ChIA-PET data (58), H3K27ac HiChIP data (59), or Hi-C (15) data (Supplemental Figure S15). Thus, these SNPs might regulate target gene expression through direct enhancer-promoter loops. One single SNP might also regulate multiple distant genes from these annotations. For example, rs7328203, rs9533100, rs35860234 and rs9594746, interact with both the *AKAP11* and *TNFSF11* promoters (Figure 5E), implying potentially that these SNPs contribute to the regulation of expression of both genes.

DISCUSSION

To achieve genome-wide high resolution of chromatin interaction maps such as 1kb and higher resolution, exhaustive sequencing is needed for current experimental strategies: in situ Hi-C (final 3.7 billion PETs) (15), its derivatives CAP-C (final 3.2 billion PETs for CAPC-BLCAPC_B01, other samples all nearly more than 1 billion PETs) (65), and Micro-C (final 2.6 billion PETs) (66). The demanding sequencing depth limits their application and analyzing the derived sequencing data cannot be easily performed with modest hardwares and limited time. In comparison, the interaction enrichment methods such as TrAC-looping and Hi-TrAC represents another profiling perspective: obtaining comprehensive interactions among accessibility sites with a modest sequencing depth (100 million informative PETs for all downstream analysis). We present cLoops2 here for com-

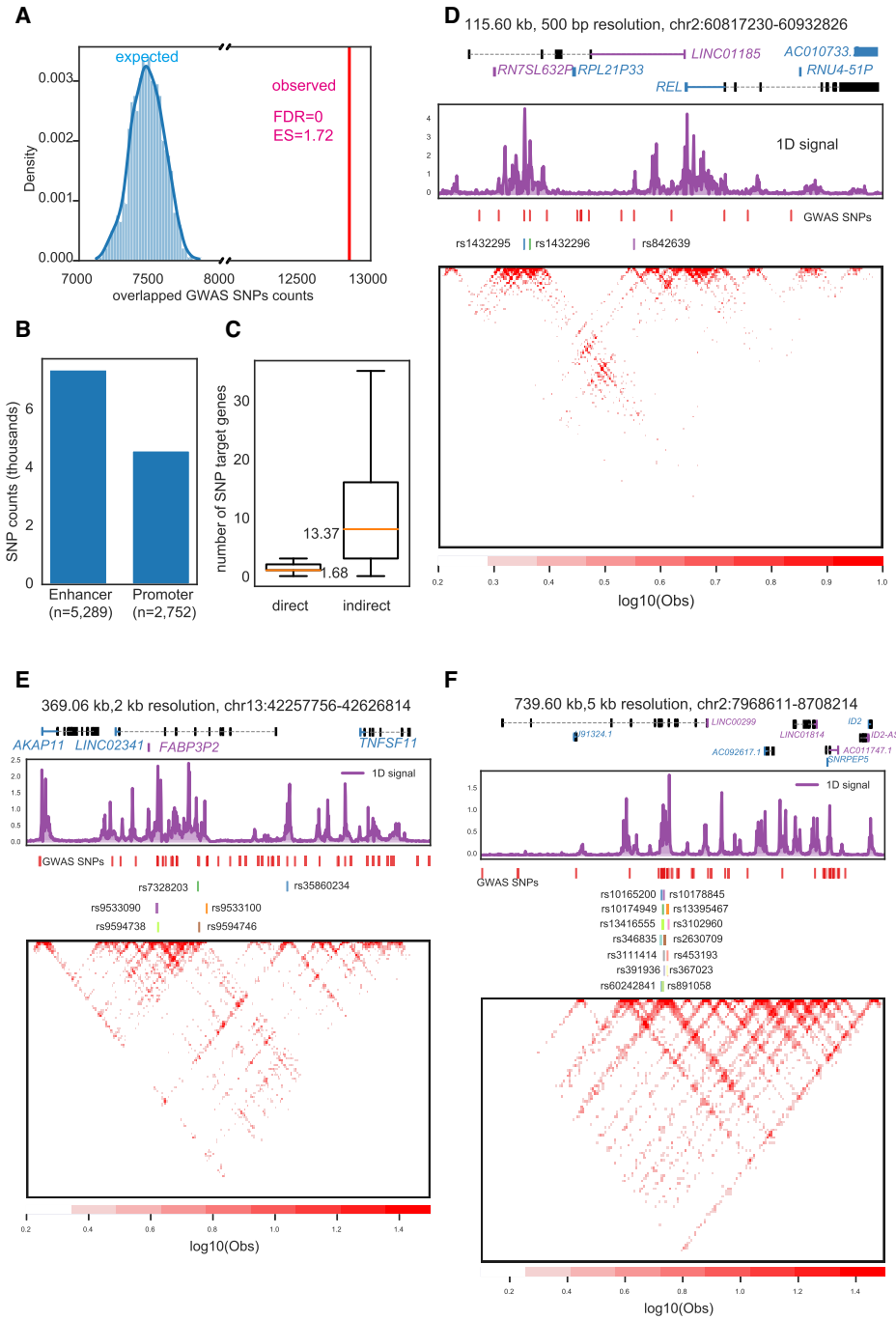

Figure 5. cLoops2 identifies SNP target genes through loops. (**A**) A large number of the GWAS cataloged SNPs overlapped with the Hi-TrAC loop anchors in GM12878 cells. Expected distribution in blue was generated from 1000 times permutation test of GWAS SNPs overlapping with sampling same chromosome, same size, and same number of genomic regions to loop anchors but without any overlaps. The red line marks the number of total overlapped GWAS SNPs with loop anchors. FDR stands for false discovery rate drawn from the permutation test, and ES stands for enrichment score of overlapped SNP number comparing to mean value of the expected distribution. (**B**) Overlapped GWAS cataloged SNPs distributions on Hi-TrAC enhancer and promoter anchors. (**C**) The direct and indirect target genes of GWAS catalogged SNPs. The target genes were annotated by cLoops2 with loops called from GM12878 Hi-TrAC data. A direct target gene referred to a gene that an SNP was located in one of the loop anchors and the other anchor of the loop was located to the promoter region of the gene. An indirect target gene refered to a gene that an SNP was located in a loop anchor, and the other anchor of the loop was not located directly to the promoter region of the gene but instead was linked to the promoter through one or more loops. (**D**) Example of SNP target gene identified by Hi-TrAC loops. The SNPs were mapped to the lincRNA gene *LINC01185*, while the Hi-TrAC data indicated that the SNPs actually targeting the *REL* gene through enhancer-promoter looping. The blue color marked the first exon of the gene in the positive strand, and the purple color marked the gene's first exon in the negative strand. SNPs colored other than red were SNPs found located in lncRNA but were looped to the promoter regions of other genes. (**E**) Example of SNP target gene identified by Hi-TrAC loops. The SNPs were mapped to the lincRNA gene *LINC02341*, while the Hi-TrAC data indicated that the SNPs are actually targeting the *AKAP11* and *TNFSF11* genes through chromatin looping. (**F**) Example of SNP target gene identified by Hi-TrAC loops. The SNPs were mapped to the lincRNA gene *LINC00299*, while the Hi-TrAC data indicated that the SNPs are actually targeting the *ID2* gene through chromatin looping.

prehensive interpretations of TrAC-looping, Hi-TrAC and similar data, especially for loop-centric analysis.

In conclusion, cLoops2 showed the general applicability of the density-based clustering algorithm blockDBSCAN for finding features in sequencing-based enriched data derived from either ChIP-seq like 1D profiling technologies or interaction data loop-centric analysis. We anticipate that cLoops2 will be a valuable tool for the broader research community, especially for similar data to TrAC-looping and Hi-TrAC.

A large number of ChIP-seq samples have been studied by ENCODE (67) and NIH Roadmap Epigenomics (68), and a large number of samples on chromatin interaction by 4DN (69) consortia and other studies are emerging (58). More challenges on analyzing these datasets will be met with the accumulation of chromatin interaction data, and we anticipate that cLoops2 may be broadly applied to addressing the challenges for the community.

DATA AVAILABILITY

Sequencing data generated in this study have been deposited to the Gene Expression Omnibus database with the accession of GSE179010. Visualization tracks are available through the WashU Epigenome Browser with session bundle id: e5405410-e917-11eb-871e-f338f56405eb. Results generated by cLoops2 in this study can be found in https://github.com/YaqiangCao/cLoops2_supp. All code of cLoops2 is open-source and available at GitHub, stable version at https://github.com/KejiZhaoLab/cLoops2and latest version at: https://github.com/YaqiangCao/cLoops2.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the NHLBI DNA sequencing facility for sequencing the NGS libraries. We thank Zhaoxiong Chen and Jing-Dong Jackie Han for the essential discussion of the blockDBSCAN algorithm and Jonathan Perrie for cLoops2 documentation and testing help. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). This research was supported by the Intramural Research Program of the National Institute of Health, National Heart, Lung and Blood Institute, NIH.

Author contributions: Y.Q.C. and K.Z. designed the project; Y.Q.C. implemented the blockDBSCAN algorithm; Y.Q.C. finished all coding, testing, documentation and analysis. S.L. and Q.T. carried out the Hi-TrAC and TrAC-looping experiments, and G.R. carried out the ChIC-seq experiments. All authors contributed to data analysis, interpretation and writing the paper.

FUNDING

Intramural Research Program of the National Heart, Lung, and Blood Institute (to K.Z.); 4DN Transformative Collaborative Project Award [A-0066 to K.Z.]. Funding for open access charge: Intramural Research Program of the National Heart, Lung, and Blood Institute; 4DN Transformative Collaborative Project Award [A-0066].

Conflict of interest statement. None declared.

REFERENCES

1. Szabo,Q., Bantignies,F. and Cavalli,G. (2019) Principles of genome folding into topologically associating domains. *Sci. Adv.*, **5**, eaaw1668.
2. Bouwman,B.A. and de Laat,W. (2015) Getting the genome in shape: the formation of loops, domains and compartments. *Genome Biol.*, **16**, 154.
3. Lupianez,D.G., Kraft,K., Heinrich,V., Krawitz,P., Brancati,F., Klopocki,E., Horn,D., Kayserili,H., Opitz,J.M., Laxova,R. *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**, 1012–1025.
4. Hnisz,D., Weintraub,A.S., Day,D.S., Valton,A.L., Bak,R.O., Li,C.H., Goldmann,J., Lajoie,B.R., Fan,Z.P., Sigova,A.A. *et al.* (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, **351**, 1454–1458.
5. Zheng,H. and Xie,W. (2019) The role of 3D genome organization in development and cell differentiation. *Nat. Rev. Mol. Cell Biol.*, **20**, 535–550.
6. Naumova,N., Imakaev,M., Fudenberg,G., Zhan,Y., Lajoie,B.R., Mirny,L.A. and Dekker,J. (2013) Organization of the mitotic chromosome. *Science*, **342**, 948–953.
7. Zhang,H., Emerson,D.J., Gilgenast,T.G., Titus,K.R., Lan,Y., Huang,P., Zhang,D., Wang,H., Keller,C.A., Giardine,B. *et al.* (2019) Chromatin structure dynamics during the mitosis-to-G1 phase transition. *Nature*, **576**, 158–162.
8. Kim,Y., Shi,Z., Zhang,H., Finkelstein,I.J. and Yu,H. (2019) Human cohesin compacts DNA by loop extrusion. *Science*, **366**, 1345–1349.
9. Davidson,I.F., Bauer,B., Goetz,D., Tang,W., Wutz,G. and Peters,J.M. (2019) DNA loop extrusion by human cohesin. *Science*, **366**, 1338–1345.
10. Vian,L., Pekowska,A., Rao,S.S.P., Kieffer-Kwon,K.R., Jung,S., Baranello,L., Huang,S.C., El Khattabi,L., Dose,M., Pruett,N. *et al.* (2018) The energetics and physiological impact of cohesin extrusion. *Cell*, **173**, 1165–1178.
11. Weintraub,A.S., Li,C.H., Zamudio,A.V., Sigova,A.A., Hannett,N.M., Day,D.S., Abraham,B.J., Cohen,M.A., Nabet,B., Buckley,D.L. *et al.* (2017) YY1 is a structural regulator of enhancer-promoter loops. *Cell*, **171**, 1573–1588.
12. Haarhuis,J.H.I., van der Weide,R.H., Blomen,V.A., Yanez-Cuna,J.O., Amendola,M., van Ruiten,M.S., Krijger,P.H.L., Teunissen,H., Medema,R.H., van Steensel,B. *et al.* (2017) The cohesin release factor WAPL restricts chromatin loop extension. *Cell*, **169**, 693–707.
13. Zhou,Q., Yu,M., Tirado-Magallanes,R., Li,B., Kong,L., Guo,M., Tan,Z.H., Lee,S., Chai,L., Numata,A. *et al.* (2021) ZNF143 mediates CTCF-bound promoter-enhancer loops required for murine hematopoietic stem and progenitor cell function. *Nat. Commun.*, **12**, 43.
14. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,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.
15. Rao,S.S.P., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
16. Tang,Z., Luo,Oscar J., Li,X., Zheng,M., Zhu,Jacqueline J., Szalaj,P., Trzaskoma,P., Magalska,A., Wlodarczyk,J., Ruszczycki,B. *et al.* (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, **163**, 1611–1627.
17. Mumbach,M.R., Rubin,A.J., Flynn,R.A., Dai,C., Khavari,P.A., Greenleaf,W.J. and Chang,H.Y. (2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods*, **13**, 919–922.
18. Fang,R., Yu,M., Li,G., Chee,S., Liu,T., Schmitt,A.D. and Ren,B. (2016) Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.*, **26**, 1345–1348.
19. Lai,B., Tang,Q., Jin,W., Hu,G., Wangsa,D., Cui,K., Stanton,B.Z., Ren,G., Ding,Y., Zhao,M. *et al.* (2018) Trac-looping measures

genome structure and chromatin accessibility. *Nat. Methods*, **15**, 741–747.
20. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
21. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
22. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
23. Liu,T., Ortiz,J.A., Taing,L., Meyer,C.A., Lee,B., Zhang,Y., Shin,H., Wong,S.S., Ma,J., Lei,Y. *et al.* (2011) Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.*, **12**, R83.
24. Ramirez,F., Ryan,D.P., Gruning,B., Bhardwaj,V., Kilpert,F., Richter,A.S., Heyne,S., Dundar,F. and Manke,T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–W165.
25. 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**, 679–685.
26. Zufferey,M., Tavernari,D., Oricchio,E. and Ciriello,G. (2018) Comparison of computational methods for the identification of topologically associating domains. *Genome Biol.*, **19**, 217.
27. Wolff,J., Rabbani,L., Gilsbach,R., Richard,G., Manke,T., Backofen,R. and Gruning,B.A. (2020) Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.*, **48**, W177–W184.
28. Harly,C., Kenney,D., Ren,G., Lai,B.B., Raabe,T., Yang,Q., Cam,M.C., Xue,H.H., Zhao,K.J. and Bhandoola,A. (2019) The transcription factor TCF-1 enforces commitment to the innate lymphoid cell lineage. *Nat. Immunol.*, **20**, 1150–1160.
29. Wang,Z., Zang,C., Cui,K., Schones,D.E., Barski,A., Peng,W. and Zhao,K. (2009) Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, **138**, 1019–1031.
30. Ren,G., Jin,W., Cui,K., Rodrigez,J., Hu,G., Zhang,Z., Larson,D.R. and Zhao,K. (2017) CTCF-mediated enhancer-promoter interaction is a critical regulator of cell-to-cell variation of gene expression. *Mol. Cell*, **67**, 1049–1058.
31. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
32. Stovner,E.B. and Saetrom,P. (2019) epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics*, **35**, 4392–4393.
33. Cao,Y.Q., Chen,Z.X., Chen,X.W., Ai,D.S., Chen,G.Y., McDermott,J., Huang,Y., Guo,X.X. and Han,J.D.J. (2020) Accurate loop calling for 3D genomic data with cLoops. *Bioinformatics*, **36**, 666–675.
34. Durand,N.C., Shamim,M.S., Machol,I., Rao,S.S., Huntley,M.H., Lander,E.S. and Aiden,E.L. (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.*, **3**, 95–98.
35. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
36. Trapnell,C., Hendrickson,D.G., Sauvageau,M., Goff,L., Rinn,J.L. and Pachter,L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, **31**, 46–53.
37. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
38. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
39. Walt,S, Colbert,S.C. and Varoquaux,G. (2011) The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.*, **13**, 22–30.
40. Durand,N.C., Robinson,J.T., Shamim,M.S., Machol,I., Mesirov,J.P., Lander,E.S. and Aiden,E.L. (2016) Juicebox provides a visualization system for Hi-C Contact maps with unlimited zoom. *Cell Syst.*, **3**, 99–101.
41. Li,D., Hsu,S., Purushotham,D., Sears,R.L. and Wang,T. (2019) WashU epigenome browser update 2019. *Nucleic. Acids. Res.*, **47**, W158–W165.
42. Saldanha,A.J. (2004) Java Treeview-extensible visualization of microarray data. *Bioinformatics*, **20**, 3246–3248.
43. Abdennur,N. and Mirny,L.A. (2020) Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*, **36**, 311–316.
44. Rowley,M.J., Poulet,A., Nichols,M.H., Bixler,B.J., Sanborn,A.L., Brouhard,E.A., Hermetz,K., Linsenbaum,H., Csankovszki,G., Lieberman Aiden,E. *et al.* (2020) Analysis of Hi-C data using SIP effectively identifies loops in organisms from C. elegans to mammals. *Genome Res.*, **30**, 447–458.
45. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
46. Meers,M.P., Tenenbaum,D. and Henikoff,S. (2019) Peak calling by sparse enrichment analysis for CUT&RUN chromatin profiling. *Epigenet. Chromatin*, **12**, 42.
47. Skene,P.J. and Henikoff,S. (2017) An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife*, **6**,e21856.
48. Ku,W.L., Nakamura,K., Gao,W., Cui,K., Hu,G., Tang,Q., Ni,B. and Zhao,K. (2019) Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. *Nat. Methods*, **16**, 323–325.
49. Phanstiel,D.H., Boyle,A.P., Heidari,N. and Snyder,M.P. (2015) Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics*, **31**, 3092–3098.
50. Lareau,C.A. and Aryee,M.J. (2018) hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. *Nat. Methods*, **15**, 155–156.
51. Bhattacharyya,S., Chandra,V., Vijayanand,P. and Ay,F. (2019) Identification of significant chromatin contacts from HiChIP data by FitHiChIP. *Nat. Commun.*, **10**, 4221.
52. Shi,C., Rattray,M. and Orozco,G. (2020) HiChIP-Peaks: a HiChIP peak calling algorithm. *Bioinformatics*, **36**, 3625–3631.
53. Zang,C., Schones,D.E., Zeng,C., Cui,K., Zhao,K. and Peng,W. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952–1958.
54. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
55. Ku,W.L., Pan,L., Cao,Y., Gao,W. and Zhao,K. (2021) Profiling single-cell histone modifications using indexing chromatin immunocleavage sequencing. *Genome Res.*, **31**, 1831–1842.
56. Goutte,C. and Gaussier,E. (2005) In: Losada,D.E. and Fernández-Luna,J.M. (eds). *Advances in Information Retrieval*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 345–359.
57. Ross,S.H. and Cantrell,D.A. (2018) Signaling and function of interleukin-2 in T lymphocytes. *Annu. Rev. Immunol.*, **36**, 411–433.
58. Grubert,F., Srivas,R., Spacek,D.V., Kasowski,M., Ruiz-Velasco,M., Sinnott-Armstrong,N., Greenside,P., Narasimha,A., Liu,Q., Geller,B. *et al.* (2020) Landscape of cohesin-mediated chromatin loops in the human genome. *Nature*, **583**, 737–743.
59. Mumbach,M.R., Satpathy,A.T., Boyle,E.A., Dai,C., Gowen,B.G., Cho,S.W., Nguyen,M.L., Rubin,A.J., Granja,J.M., Kazane,K.R. *et al.* (2017) Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.*, **49**, 1602–1612.
60. Juric,I., Yu,M., Abnousi,A., Raviram,R., Fang,R., Zhao,Y., Zhang,Y., Qiu,Y., Yang,Y., Li,Y. *et al.* (2019) MAPS: model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLoS Comput. Biol.*, **15**, e1006982.
61. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9362–9367.
62. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E.

et al. (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic. Acids. Res.*, **47**, D1005–D1012.
63. Gotoda,T. (2015) From association to function in the post-GWAS Era. *J. Atheroscler. Thromb.*, **22**, 442–444.
64. Gallagher,M.D. and Chen-Plotkin,A.S. (2018) The post-GWAS era: from association to function. *Am. J. Hum. Genet.*, **102**, 717–730.
65. You,Q., Cheng,A.Y., Gu,X., Harada,B.T., Yu,M., Wu,T., Ren,B., Ouyang,Z. and He,C. (2021) Direct DNA crosslinking with CAP-C uncovers transcription-dependent chromatin organization at high resolution. *Nat. Biotechnol.* **39**, 225–235.
66. Hsieh,T.S., Cattoglio,C., Slobodyanyuk,E., Hansen,A.S., Rando,O.J., Tjian,R. and Darzacq,X. (2020) Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Mol. Cell*, **78**, 539–553.
67. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
68. Roadmap Epigenomics,C., Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
69. Dekker,J., Belmont,A.S., Guttman,M., Leshyk,V.O., Lis,J.T., Lomvardas,S., Mirny,L.A., O'Shea,C.C., Park,P.J., Ren,B *et al.* (2017) The 4D nucleome project. *Nature*, **549**, 219–226.