Molecules and Cells

HiCORE: Hi-C Analysis for Identification of Core Chromatin Looping Regions with Higher Resolution

Hongwoo Lee[1] and Pil Joon Seo[1,2,*]

[1]Department of Chemistry, Seoul National University, Seoul 08826, Korea, [2]Plant Genomics and Breeding Institute, Seoul National University, Seoul 08826, Korea
*Correspondence: pjseo1@snu.ac.kr
https://doi.org/10.14348/molcells.2021.0014
www.molcells.org

Genome-wide chromosome conformation capture (3C)-based high-throughput sequencing (Hi-C) has enabled identification of genome-wide chromatin loops. Because the Hi-C map with restriction fragment resolution is intrinsically associated with sparsity and stochastic noise, Hi-C data are usually binned at particular intervals; however, the binning method has limited reliability, especially at high resolution. Here, we describe a new method called HiCORE, which provides simple pipelines and algorithms to overcome the limitations of single-layered binning and predict core chromatin regions with three-dimensional physical interactions. In this approach, multiple layers of binning with slightly shifted genome coverage are generated, and interacting bins at each layer are integrated to infer narrower regions of chromatin interactions. HiCORE predicts chromatin looping regions with higher resolution, both in human and *Arabidopsis* genomes, and contributes to the identification of the precise positions of potential genomic elements in an unbiased manner.

Keywords: 3D genome, binning method, CCCTC-binding factor, chromatin loop, CTCF, Hi-C

INTRODUCTION

Genome-wide chromosome conformation capture (3C)-based high-throughput sequencing (Hi-C) analysis maps all possible genomic interactions in a manner dependent on three-dimensional (3D) distance. In this method, cross-linked chromatin is digested into fragments by a restriction enzyme (Dekker et al., 2002). Then, restriction fragments are proximity-ligated to generate a library of chimeric circular DNA. Paired-end sequencing and mapping of reads to a reference genome identifies interacting restriction fragments and measures their interaction frequencies (Lieberman-Aiden et al., 2009). Unfortunately, despite the high sequencing depth of typical Hi-C experiments, single fragment–resolution contact matrices are still too sparse to analyze. Hence, to ensure efficient and reliable data processing, Hi-C data are usually binned with fixed intervals at relatively low resolution (>5 kb) (Choi et al., 2020; Fernandez-Albert et al., 2019; Lajoie et al., 2015).

Because the resolution of the Hi-C map is a critical issue for 3D genome conformation studies, multiple studies have focused on improving the resolution of raw Hi-C matrices (Cameron et al., 2020; Liu and Wang, 2019; Liu et al., 2019; Zhang et al., 2018). Although many cutting edge technologies have been implemented, including machine learning-based computational methods, the binning strategy is still an important determinant of Hi-C resolution. Two major types of binning methods have been used in practice: fixed interval binning and fragment unit binning. Fixed interval binning allows intervals to have a regular size (e.g., 5 kb) without considering restriction fragment size, whereas fragment unit binning has an interval defined by the size of restriction

Received 18 January, 2021; revised 13 October, 2021; accepted 13 October, 2021; published online 25 December, 2021

eISSN: 0219-1032

fragment(s) (Lajoie et al., 2015). Although fixed interval binning is a practical approach that facilitates efficient analysis of Hi-C data, fixed intervals mismatch restriction fragments of variable sizes, especially at a high resolution, creating bias in chromatin loop detection (Lajoie et al., 2015). Thus, fragment unit binning methods have also been improved in parallel for high-resolution Hi-C analysis. Because single fragment resolution analysis is limited due to contact matrix sparsity, as an alternative method, several restriction fragments are assembled to generate a single bin (here called multi-fragment binning) (Cameron et al., 2020; Liu et al., 2016).

Following the binning step, genomic looping regions are identified by various methods, such as Fit-HiC2 (Kaul et al., 2020), HiCCUPS (Durand et al., 2016), and cLoop (Cao et al., 2020). In particular, Fit-HiC2 is a recently updated python package that computes statistical confidence estimates for Hi-C contact maps and has been widely used for a high resolution (~1 kb) Hi-C data analysis (Kaul et al., 2020). Despite the advance in the chromatin loop detection methods, since high resolution Hi-C data analysis frequently produces noise outputs, it requires a more robust post-analysis to determine the reliability of chromatin loop prediction.

In this study, we developed the HiCORE algorithm, which includes an advanced binning method and loop prediction strategy compatible with conventional loop calling methods. HiCORE predicts fine-scale interactions between genomic elements. Independently of the binning method, single-layered binning (conventional method) is insufficient for defining chromatin looping regions, especially at high resolution. To further improve resolution and precision rate of detecting 3D chromatin interactions, HiCORE generates multiple layers of bin arrays with different genome coverages. Significantly interacting chromatin regions at each layer are identified (Kaul et al., 2020), and then this information from multiple layers is integrated to draw partially overlapping and narrower chromatin regions with 3D interactions, which represent chromatin loops with higher resolution. Prediction of high-resolution chromatin loops allows the estimation of precise positions of genomic elements.

MATERIALS AND METHODS

Hi-C data and their pre-processing

For human genome analysis, we downloaded the processed Hi-C data (mapq > 30) of chromosome 20 of GM12878 human cell line (Rao et al., 2014) by Juicer dump command (Durand et al., 2016). The high resolution Hi-C data of *Arabidopsis* were obtained from National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession number SRP032990 and SRP064711. The 4 replicates were mapped to *Arabidopsis* genome (TAIR10) and further processed by Juicer packages (Durand et al., 2016). We extracted the raw interaction matrix from the '.hic' file that was derived from reads with mapping quality (mapq) > 30, and matrix resolution was calculated as previously described (Rao et al., 2014).

Binning strategy

Fragments with a size below a given cutoff length are merged with neighboring fragments. The merged fragments are assembled into a single bin (multi-fragment bin). This binning strategy is applied to all fragments except for the end fragment. For 'forward binning', from the 5' end of the chromosome, a restriction fragment shorter than the cutoff length is merged with the following fragment (in the 3'-direction) until the merged fragment is longer than cutoff length. For 'reverse binning', from the 3' end of the chromosome, a restriction fragment shorter than the cutoff length is merged with the preceding fragment (in the 5'-direction) until the merged fragment is longer than the cutoff length. For 'random binning', HiCORE randomly selects restriction fragments for construction of a binning layer. Randomly selected fragments shorter than the cutoff length are first merged with the following fragment in the 3'-direction. In the middle of fragment assembly, if the next fragment in the 3'-position was already merged with other fragments, the bin being generated is merged with the fragment in the 5'-direction until the bin being generated is longer than the cutoff length. If the neighboring fragments in both directions are already taken by other completed bins, the bin being generated is divided into two units. The first unit (5'-neighboring fragments of the random fragment) and the second unit (the random fragment plus its 3'-neighboring fragments) of the bin being generated are separately merged to already-completed neighboring bins in the 5' and 3' positions, respectively.

Single-layered Fit-HiC2 analysis

From the extracted raw fragment matrix file, we generated '.interaction.gz' and 'fragments.gz' files that are necessary for subsequent Fit-HiC2 analysis. We calculated normalization vector of the raw matrix using 'HiCKRy.py' in Fit-HiC2 packages (Kaul et al., 2020). To identify chromatin loops at each layer, Fit-HiC2 analysis was performed with the following parameters; no fixed interval (-r 0), only intrachromosomal contacts (-x intraOnly), various looping distance thresholds (lower bound [-L] 20000 upper bound [-U] 250000, -L 250000 -U 500000, -L 500000 -U 1000000, -L 20000 -U 1000000) for human genome analysis and -r 0 -L 2000 -U 20000 -x intraOnly for *Arabidopsis* genome analysis. The output loops were filtered by a statistical cutoff value ($q < 0.01$) that indicates statistical significances of chromatin contact probability with respect to genomic distance for further analysis.

Profiling of core chromatin looping regions using HiCORE

The raw fragment matrix file was assigned to each bin layer, and then single-layered Fit-HiC2 analysis was performed in each bin layer. After defining the interacting chromatin regions using Fit-HiC2 packages, a statistical cutoff value ($q < 0.01$) is used to obtain statistically significant anchor regions. Fit-HiC2 output files, which contain 3D interaction data in units of multi-fragment bins, are converted to data in units of single-fragment bins for each binning layer. Then, the output files are integrated, and the layer-detection frequency of each single fragment having 3D interactions from total binning layers (No. of detecting layers/No. of total layers) was calculated by HiCORE. Restriction fragment(s) having local maximum values of the layer-detection frequency above the certain threshold were specified as core looping regions. The

size of anchor regions was calculated by square root value of the identified looping area.

Aggregate peak analysis

Aggregate peak analysis (APA) was performed at restriction fragment unit. The 20 × 20 fragment pairs centered with restriction fragment pairs having 3D interactions were aggregated by in-house script.

CCCTC-binding factor motif density analysis

The CCCTC-binding factor (CTCF) motif sites in human genome were collected by fimo in MEME suite with motif ID MA0139.1. The CTCF motif density was calculated by (the number of CTCF motifs in total identified anchor regions)/(total size [bp] of identified anchor regions). For the analysis, HiCORE integrated 80 binning layers with a threshold bin size of 1,000 bp.

Loop recovery rate and false discovery rate analyses

Reference chromatin loops at the 5 kb resolution for GM12878 cell line were obtained from public datasets (see Dataset description). Fit-HiC2 (Kaul et al., 2020), SIP (Rowley et al., 2020), and Mustache (Roayaei Ardakany et al., 2020) were performed with 1 kb resolution binning, whereas HiCORE was performed with both 2 kb and 5 kb bin size thresholds for each binning layer, which define anchor regions with the final size comparable to 1 kb fixed interval binning. Statistical cutoff ($q < 0.01$ in Fit-HiC2, $q < 0.1$ in SIP, and $p < 0.1$ in Mustache) was applied to obtain the reasonable number of chromatin loops. Since HiCORE usually identifies a large number of chromatin loops, we selected the top 1,000 most statistically significant loops at each binning layer and performed HiCORE analysis with 0.2 cutoff value of layer-detection frequency. Looping distances ranging from 20 to 1,000 kb were collected for the comparison with various loop calling methods. Identified chromatin loops were sorted by statistical significances (Fit-HiC2, SIP, and Mustache) or layer-detection frequency (HiCORE). The fraction of recovered reference loops was calculated by the number of reference loops overlapped with identified loops divided by the total number of reference loops. False discovery rate (FDR) was calculated by the number of identified loops that are not overlapped with the reference loops divided by the total number of identified loops.

Reproducibility analysis

Reproducibility of HiCORE was measured by anchor recovery rate of HiCORE-identified loops. The anchor recovery rate was calculated by square root value of (overlapped looping regions between two replicates divided by total 3D chromatin looping regions of a single replicate). To generate downsampled technical replicates, mapped read pairs were obtained from GEO accession GSE63525 (Rao et al., 2014), and 50%, 70%, or 80% of total mapped read pairs (mapq > 30) were randomly sampled for downstream analysis.

RESULTS AND DISCUSSION

Framework of HiCORE

HiCORE generated multiple layers of bin arrays, in which a single bin was defined through serial assembly of fragments (multi-fragment binning) to have a size above a certain threshold. The first layer was generated by forward multi-fragment binning initiated from the starting point (5'-end) of the genome (forward binning), and the second layer was created by reverse multi-fragment binning initiated from the end point (3'-end) of the genome (reverse binning). Additional layers were generated by bidirectional binning from randomly selected positions of the genome (random binning). Bin arrays at each layer could be shifted a bit relative to one another, and thus have different genome coverages (Fig. 1). An interaction-frequency matrix of fragment–resolution Hi-C data was assigned to bin arrays at each layer, and bin pairs with statistically significant interactions were identified using the Fit-HiC2 package (Kaul et al., 2020). Identified bins having 3D interactions in all layers are integrated by HiCORE

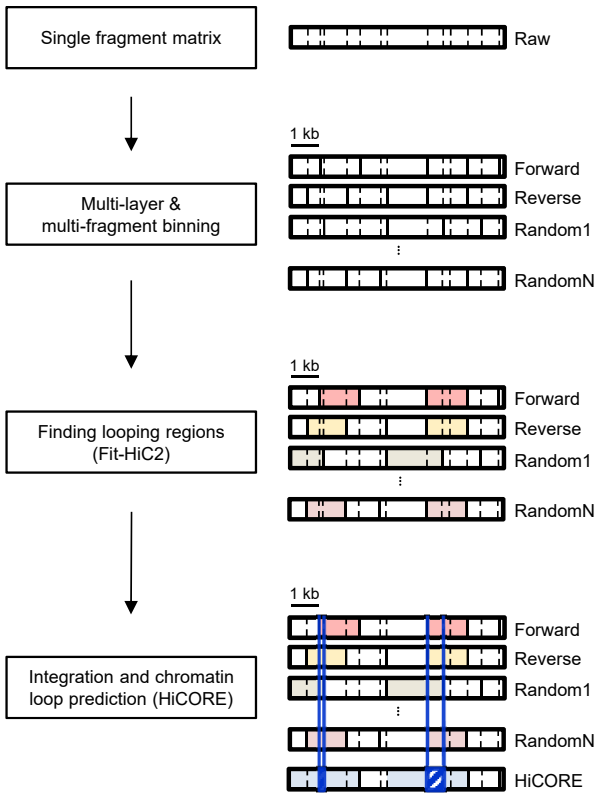

Fig. 1. HiCORE features. Schematic diagram showing the working process of HiCORE. Multiple layers of multi-fragment binning are produced, and chromatin looping regions at each layer are predicted based on raw 3D contact matrices. HiCORE identifies overlapped regions of interacting bins from multiple layers, specifying anchor regions with 3D interaction at higher resolution. Dashed lines indicate restriction sites. Solid lines indicate multi-fragment bin units larger than a given cutoff bin size (scale bars). Colored regions in each layer indicate chromatin looping regions. HiCORE suggests core interacting regions (hatched regions), based on the local maximum values of the layer-detection frequency.

(light blue-colored regions in Fig. 1), and all chromatin regions with potential 3D interactions were collected in an unbiased manner. HiCORE further calculated layer-detection frequency of each single fragment from total binning layers (No. of detecting layers/No. of total layers), and predicted core chromatin looping regions by specifying local maximum values of the layer-detection frequency (blue-hatched regions in Fig. 1). In this study, we performed HiCORE analysis after Fit-HiC2 application, but HiCORE can be also applied to other conventional chromatin loop calling tools, such HiCCUPS, SIP, and MUSTACHE, in the place of Fit-HiC2, indicating that HiCORE can be used to further specify the chromatin looping regions.

Dataset description

We intended to develop a precise loop prediction tool independently of TAD structures because some species with relatively small genome had no clear TAD structures. As a research model for the study, *Arabidopsis* genome has no clear 3D chromatin conformation as well as TAD structures (Feng et al., 2014; Wang et al., 2015). Moreover, genes encoding CTCF insulator proteins have not been identified in the *Arabidopsis* genome (Ong and Corces, 2009). However, several studies raised the possibility that chromatin looping

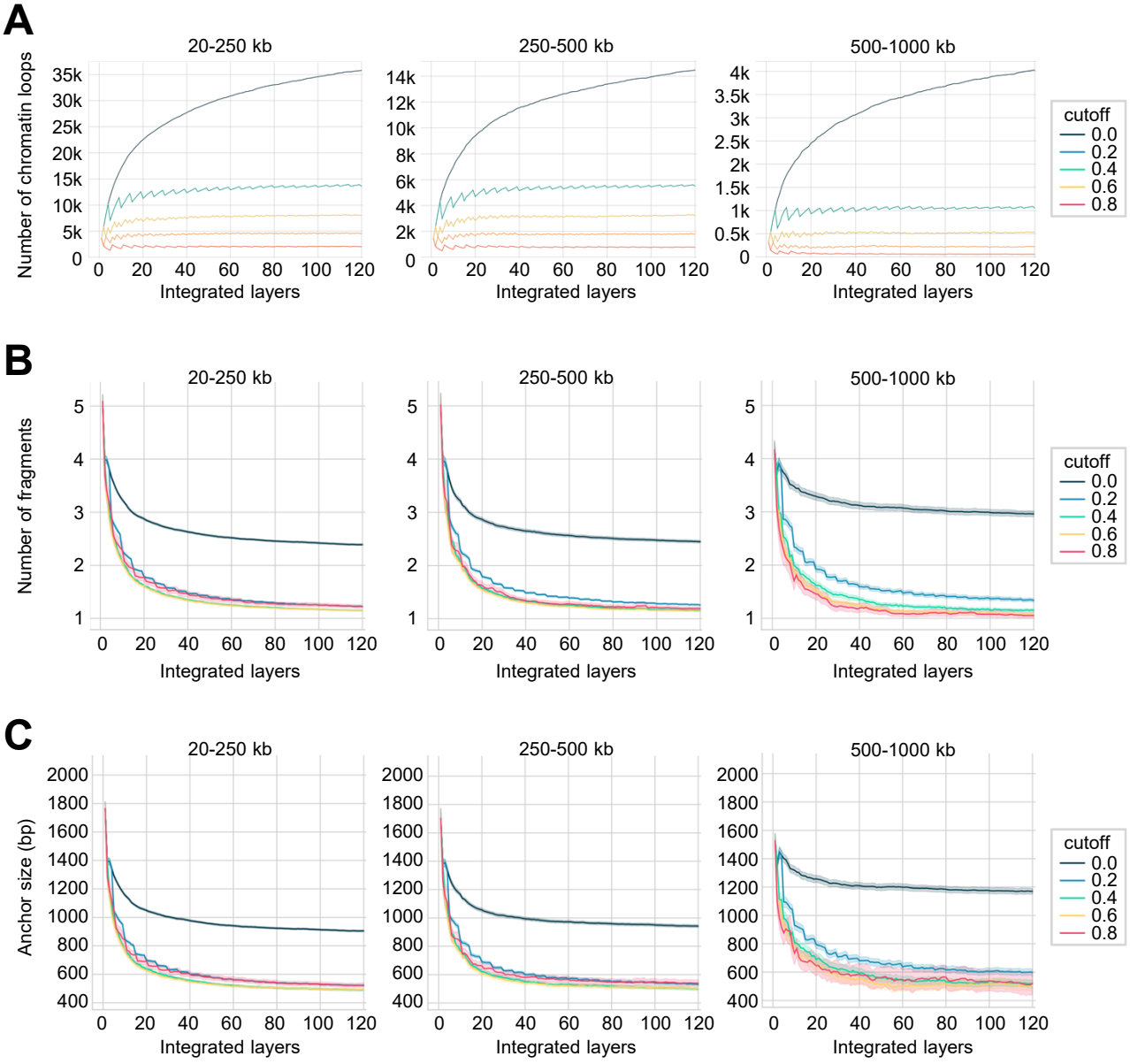

Fig. 2. HiCORE test results for specification of anchor regions. (A) The number of identified chromatin loops using HiCORE. (B) The average number of fragments in each specified anchor regions. (C) The average size (bp) of specified anchor regions. One bin layer was added at a time, repeatedly, and anchor regions were specified at various ranges of looping distances (shown above each graph). The x-axis indicates the number of layers integrated for HiCORE analysis. Cutoff values indicate the layer-detection frequency.

is crucial for gene regulation in *Arabidopsis* (Crevillén et al., 2013; Liu et al., 2016; Sun et al., 2020). Hence, we wanted to implement a stringent method for 3D genome analysis that identifies reliable chromatin looping regions in the model plant genome. To validate the logic of our algorithms, we applied HiCORE to human Hi-C analysis (Rao et al., 2014) and compared with conventional single-layered binning methods as well as various loop calling methods. For the comparative analysis, reference chromatin loops were obtained from public datasets, including HiCCUPS-identified Hi-C loops (Durand et al., 2016; Rao et al., 2014), Cohesin and H3K27ac HiChIP (Mumbach et al., 2016; 2017), RAD21 and CTCF ChIA-PET (GSM1436265, Heidari et al., 2014; GSM1872886, Tang et al., 2015), CTCF, RAD21, SMC, and H3K27ac ChIP-seq datasets (ENCFF797LSC, ENCFF416LQD, ENCFF416LQD, and ENCFF440GZA).

Specification of chromatin looping regions by HiCORE

To validate the relevance of HiCORE in specifying chromatin looping regions from human Hi-C data (Rao et al., 2014), we generated multiple layers of bin arrays, in which a single bin was defined through multi-fragment binning to have a size above 1,000 bp. The threshold bin size was determined based on the matrix resolution of the original human Hi-C data (Rao et al., 2014). We integrated information about 3D chromatin interactions from multiple layers and identified all possible chromatin loops. The number of identified chromatin loops were gradually increased during HiCORE analysis without cutoff value of layer-detection frequency, because different loops can be identified from different binning layers, which were cumulative as binning layers are integrated (Fig. 2A). However, at particular cutoff values of layer-detection frequency, the number of chromatin loops was relatively low and saturated after ~40-layer integration (Fig. 2A).

Then, we analyzed the relevance of the chromatin loops identified by HiCORE. Analysis with a single layer of binning (conventional method; integrated layer = 1) revealed that the average size of all specified anchor regions with 3D contacts was about 4 to 5 fragments at all ranges of looping distance (Fig. 2B). However, the HiCORE-generated multiple layers enabled further specification of chromatin looping regions. Compared with single-layered binning, the average size of all specified anchor regions with 3D contacts was reduced by ~80% (1.1 to 1.4 fragments) after integration of 80 binning layers (Fig. 2B), regardless of looping distance, demonstrating that anchor regions for chromatin loops are better specified after HiCORE analysis (Figs. 2B and 2C). Surprisingly, when 80 layers of bin arrays were overlapped, 72% of all predicted chromatin regions with 3D interactions were specified at single-fragment resolution with cutoff value of 0.4 (Supplementary Fig. S1). We further tested our HiCORE algorithm with various threshold of minimum bin size (400 to 1,200 bp) and confirmed that anchor regions could be further specified by HiCORE in all threshold size examined (Supplementary Fig. S2). Moreover, APA showed that HiCORE better specified core interacting fragments, which had strong 3D contacts at a single fragment center (Fig. 3, Supplementary Fig. S3).

HiCORE could considerably specify chromatin interacting regions at single-fragment resolution, but it is notable that HiCORE results could not be obtained by single-layered sin-

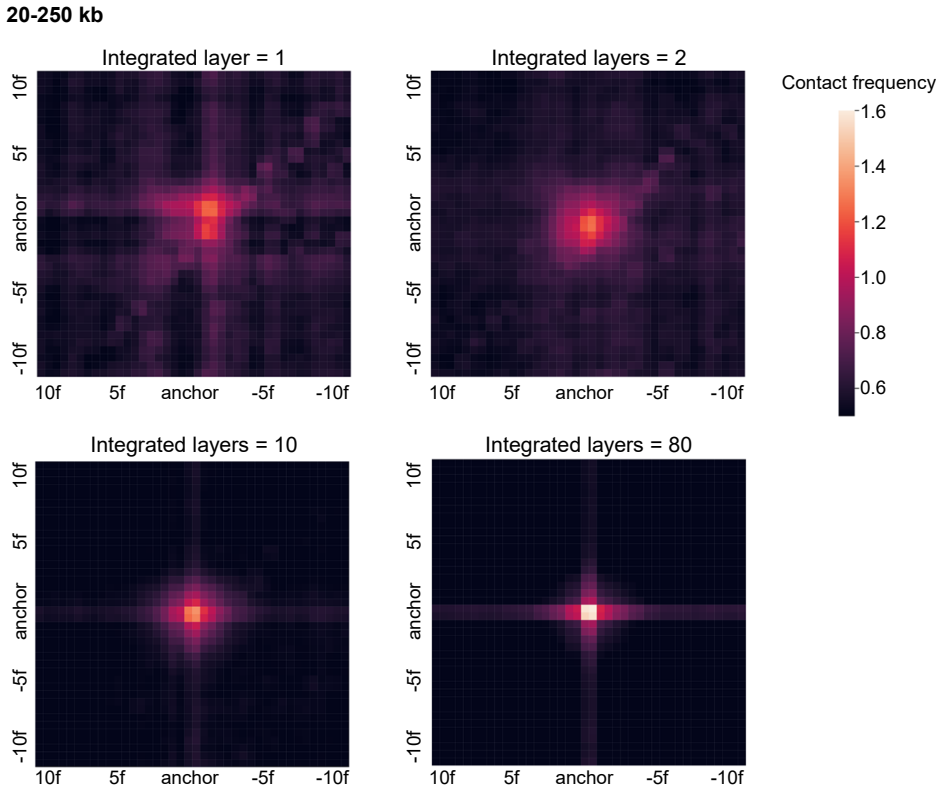

Fig. 3. Aggregate peak analysis (APA) of HiCORE-specified loops. HiCORE-specified loops at ranges of looping distance of 20-250 kb were analyzed by APA. The specified anchor fragments were center-located, and their flanking 20 fragments were also shown. A square unit indicates 0.5 fragment in the Hi-C heatmap.

gle fragment binning. The average size of HiCORE-specified anchor regions with 3D contacts was much smaller than that by the conventional single-layered Fit-HiC2 analysis at single fragment resolution (Supplementary Fig. S4). Furthermore, in terms of the number of chromatin loops, single-layered single fragment analysis produced biased results that were highly dependent on looping distances, while both single-layered fixed-interval Fit-HiC2 and HiCORE analyses could predict chromatin loops with lower bias in looping distances (Supplementary Fig. S5). It seems likely that while 3D contact information for small restriction fragments was extremely sparse at single fragment resolution, HiCORE could overcome at least in part the sparsity issue.

We also showed the results in units of base-pair length (Supplementary Fig. S6). However, at extremely high resolution (cutoff value < matrix resolution), the average size of all specified anchor regions with 3D contacts in units of base-pair length did not gradually decrease as binning layers were integrated (Supplementary Fig. S6), although the average fragment number was reduced (Supplementary Fig. S2). The inconsistency was possibly due to the sparsity of 3D contact matrices in smaller bins, which caused low reproducibility at each layer and thereby are excluded during HiCORE analysis. Thus, HiCORE can specify the core chromatin anchor regions at higher resolution, but optimal analysis can usually be performed above the matrix resolution.

Advantages of HiCORE in prediction of chromatin looping regions

To examine the biological relevance of our algorithm, we asked whether HiCORE specify the core interacting fragments that significantly overlap with CTCF-binding motifs. As more binning layers are integrated, a portion of restriction fragments with CTCF motifs to total restriction fragments in defined anchor regions was gradually increased (Fig. 4A), which is more effective at short-range interactions (Fig. 4A, Supplementary Fig. S7), indicating that CTCF motifs are enriched in HiCORE-specified anchor regions. We also compared HiCORE with a conventional single-layered binning method. HiCORE-specified anchor regions showed slightly higher CTCF-motif density (No. of CTCF motifs/size [bp] of specified anchor regions) (Fig. 4B).

To support the strength of our algorithm, we further compared HiCORE with conventional loop calling methods, such as Fit-HiC2 (Kaul et al., 2020), HiCCUPS (Durand et al., 2016), SIP (Rowley et al., 2020), and Mustache (Roayaei Ardakany et al., 2020), in detecting the reference chromatin loops at high resolution (≤1 kb) (Heidari et al., 2014; Mumbach et al., 2016; 2017; Rao et al., 2014; Tang et al., 2015). The different number of chromatin loops having looping distance ranging from 20 kb to 1,000 kb was identified depending on the methods: 1,701 loops in Fit-HiC2, 265 loops in SIP, and 697 loops in Mustache. We excluded HiCCUPS in comparison because HiCCUPS did not identify any chromatin loop at 1 kb resolution. HiCORE was performed with 2 kb and 5 kb bin size thresholds at each binning layer, and HiCORE-specified resulting chromatin loops mostly had higher resolution than 1 kb in both size thresholds (Supplementary Fig. S8). Since HiCORE usually identified a large number of chromatin loops (Fig. 2A), we collected only top 1,000 most statistically significant chromatin loops from each binning layer for the HiCORE analysis. After integrating 80 binning layers each with 1,000 chromatin loops, 1,888 and 2,102 chromatin loops were finally identified by HiCORE with 2 kb and 5 kb bin size thresholds (0.2 cutoff value of layer-detection frequency), respectively (Supplementary Fig. S8). Recovery rate analysis of reference chromatin loops revealed that HiCORE showed outstanding recovery rate compared with other loop calling methods, in spite of more specified anchor regions (Fig. 5). In addition, HiCORE showed relatively low FDR than SIP and Mustache (Supplementary Fig. S9), while Fit-HiC2 analysis showed the lowest FDR (Fig. 5, Supplementary Fig. S9). It was noteworthy that HiCORE-specified anchor regions were precisely matched with CTCF motifs and CTCF, RAD21, SMC3, and H3K27ac ChIP-seq profiles, compared with other loop calling methods, in several chromatin looping regions

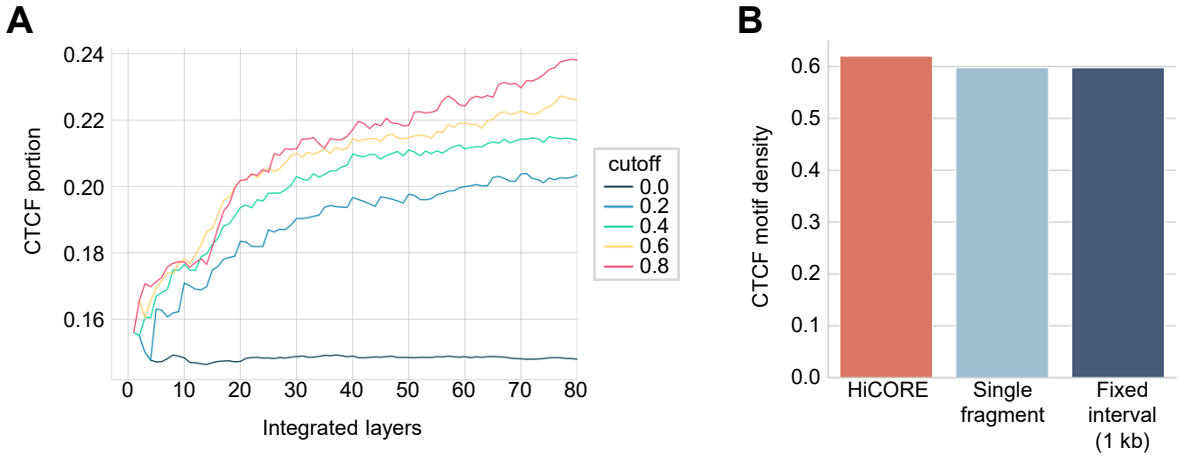

Fig. 4. Coverage of CTCF-binding sites. (A) Portion of HiCORE-specified anchor fragments with CTCF motifs to total identified fragments. The cutoff values indicate the layer-detection frequency. (B) Comparison between conventional and HiCORE analyses in specification of CTCF-containing anchor regions. The CTCF motif density was calculated by (the number of CTCF motifs/total size [kb] of identified anchors).

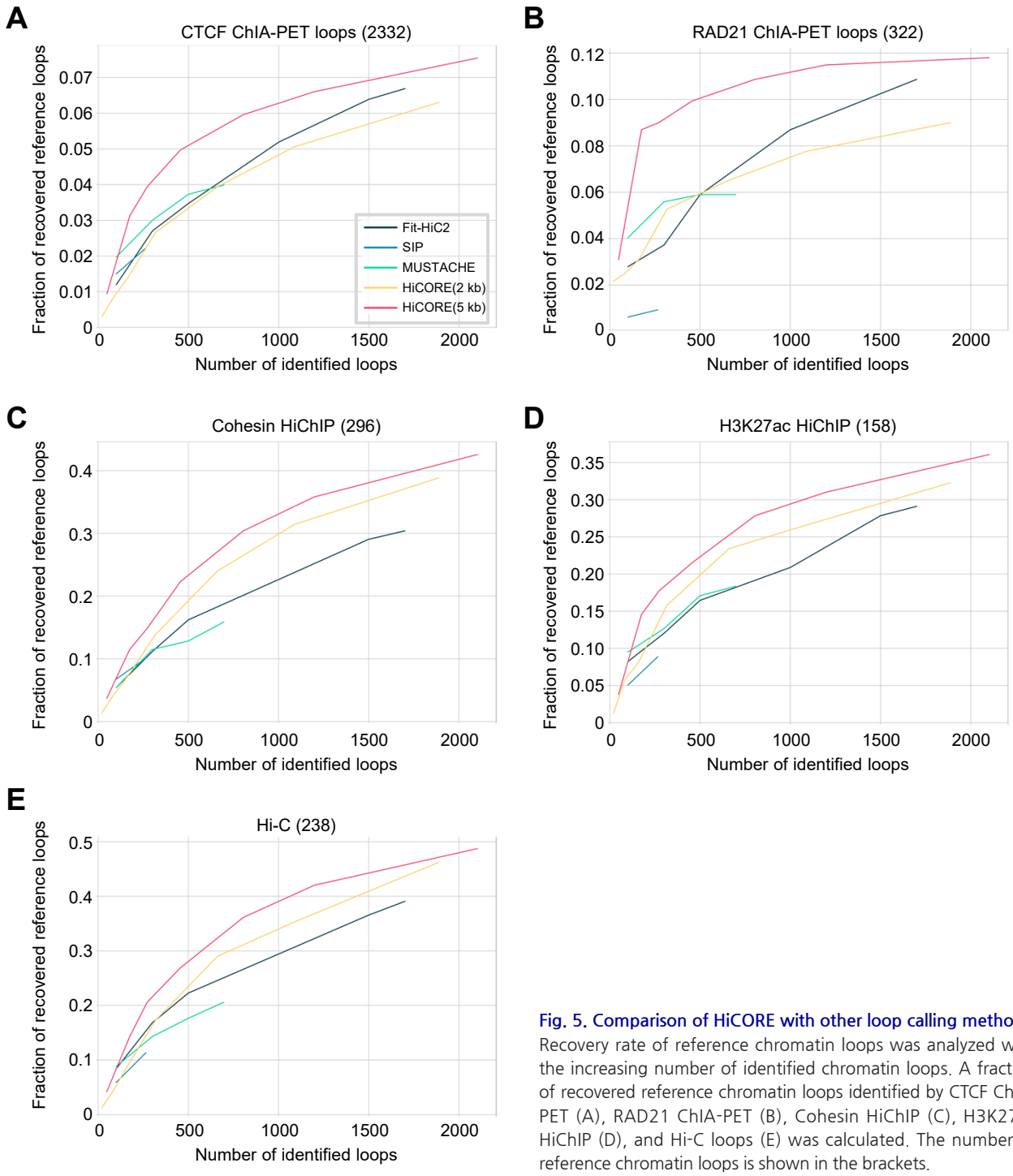

Fig. 5. Comparison of HiCORE with other loop calling methods. Recovery rate of reference chromatin loops was analyzed with the increasing number of identified chromatin loops. A fraction of recovered reference chromatin loops identified by CTCF ChIA-PET (A), RAD21 ChIA-PET (B), Cohesin HiChIP (C), H3K27ac HiChIP (D), and Hi-C loops (E) was calculated. The number of reference chromatin loops is shown in the brackets.

(Supplementary Fig. S10), indicating that HiCORE specifies high-resolution chromatin loops, which cannot be obtained by conventional loop calling methods. The performance of HiCORE was highly relevant in specifying chromatin loops particularly at high resolution (≤1 kb), and conventional methods are likely sufficient to identify chromatin loops at relatively low resolution (≥5 kb). Therefore, HiCORE can be implemented with any loop calling method to further specify chromatin looping regions, when conventional loop calling methods do not produce results with sufficient resolution.

To validate the reproducibility of our algorithm, we first performed two independent HiCORE analyses (80-layer integration) for same dataset. The two independent HiCORE analyses generated the almost same number of loops with comparable anchor size, regardless of cutoff values (Figs. 6A and 6B). At cutoff value of 0.8, ~80% of anchor regions

were reproducibly predicted (Fig. 6C) and showed similar specification of the chromatin regions overlapping with CTCF motifs each other (Fig. 6D). We also examined the reproducibility using technical replicates, which are produced by random downsampling (50%, 70%, or 80%) of total raw mapped reads from the GM12878 human cell line Hi-C data (Rao et al., 2014). Notably, HiCORE-identified loops showed higher reproducibility, when 80 layers were integrated, compared with single-layered Fit-HiC2 analysis (Fig. 6E). Especially, HiCORE showed higher performance for 50%-downsampled data and long-range (500-1,000 kb) data, which have higher sparsity in 3D chromatin interactions (Fig. 6E, Supplementary File S1). These results suggest that HiCORE is relevant in chromatin loop prediction.

However, the HiCORE algorithm also has several limitations. Since this analysis put a basis on the randomness of binning, HiCORE sometimes has low reproducibility. This flaw is exaggerated if random binning layers are not saturated to cover all possible binning combination. Indeed, independent random binning for replicates lowered reproducibility (Supplementary File S1). Furthermore, the trade-off between chromatin looping region specification and reproducibility was also observed, as shown by higher reproducibility of HiCORE analysis integrating the low number of binning layers (e.g., n = 20) (Fig. 6E, Supplementary File S1), which is due to an intrinsically low probability to overlap between highly specified, small size anchors as well as high variability of downsampled Hi-C matrices at high resolution. Overall, although there are still limitations for high resolution Hi-C analysis, HiCORE provides a new chance to improve high resolution 3D genome analysis with novel binning method. More effective and efficient computational analyses, along with technical advances in 3D genome interaction sequencing, might be necessary to extensively understand 3D genome

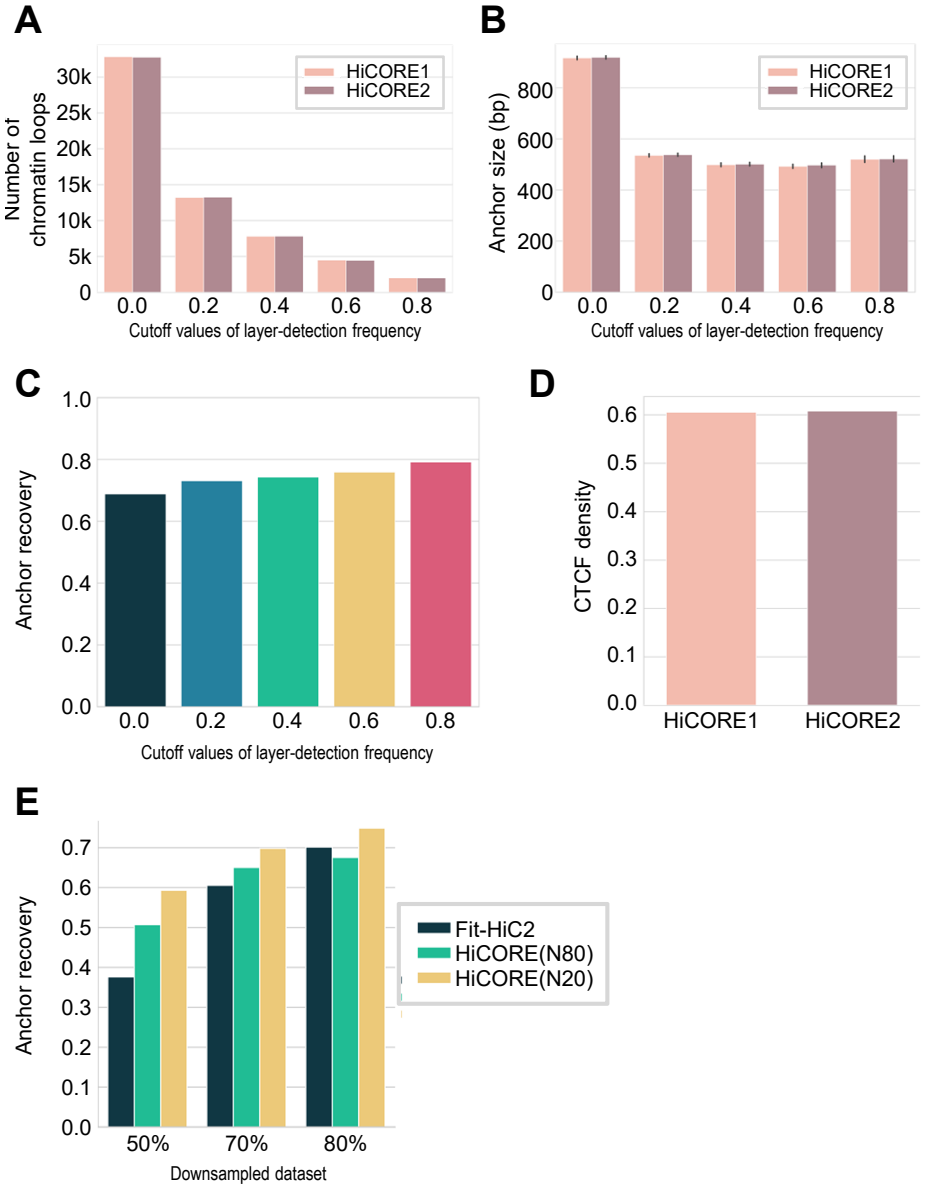

Fig. 6. Reproducibility of HiCORE analysis. (A) The number of chromatin loops. (B) The average size (bp) of specified anchor regions. (C) Anchor recovery rates of two HiCORE analyses. (D) CTCF motif density analysis. The CTCF motif density was shown as (the number of CTCF motifs/1 kb anchor regions). (A-D) Two independent HiCORE analyses were performed with the integration of 80 binning layers. The cutoff values indicate the layer-detection frequency. (E) Reproducibility of HiCORE results and 1 kb-fixed interval Fit-HiC2 analysis for downsampled datasets. HiCORE analysis (2 kb bin size threshold) was performed by integrating 80 binning layers (N80) and 20 binning layers (N20). Equal random binning layers were used for two downsampled datasets.

structure at high resolution in the future.

Applications of HiCORE to *Arabidopsis* genome

Our algorithm can also be applied to other smaller genomes, including *Arabidopsis* genome. Indeed, the HiCORE analysis significantly reduced the size (bp) and average fragment number of all specified anchors regions in *Arabidopsis* genome, compared with a conventional single layer of binning (integrated layer = 1) (Fig. 7A).

Notably, a chromatin loop in the *Arabidopsis FLOWERING LOCUS C* (*FLC*) locus, which was empirically validated in several studies (Crevillén et al., 2013; Liu et al., 2016), was not predicted by all conventional binning methods (fixed 1 kb interval binning, single fragment binning, and forward multi-fragment binning). However, HiCORE precisely identified a chromatin loop in the *FLC* locus at cutoff value of 0.25 (Fig. 7B), which corresponded to the empirically-proven loop (Crevillén et al., 2013). Considering that small chromatin loops (few kb scales) are pervasive in *Arabidopsis* genome, HiCORE will greatly contribute to find core genomic elements for chromatin looping at higher resolution in future studies of *Arabidopsis* genome.

Hi-C is a commonly used technique for mapping 3D chromatin organization. Despite significant advances in the accuracy and resolution of genome contacts, identification of fine-scale local chromatin loops remains challenging, mostly because of low sequencing depth and single-layered binning. The HiCORE algorithm takes advantage of an advanced binning method, and also predicts higher-resolution chromatin looping by generating multiple layers of bin arrays, which are slightly shifted relative to each other. Extraction of partially overlapping, more narrowly interacting regions allows higher-resolution analysis of chromatin looping from Hi-C data which most likely contain key genomic elements for 3D chromatin interaction. Furthermore, HiCORE provides an opportunity to identify chromatin loops that cannot be detected by existing binning methods. However, it is currently difficult to validate whether the HiCORE-specified regions are truly meaningful for chromatin contacts, because there are no reference studies to show empirically-proven chromatin loops at single fragment resolution. Although we validated our strategies by analyzing the density of CTCF-binding DNA motifs in specified anchor regions, it also has a limitation in that the DNA motifs do not always involve chromatin looping. Overall,

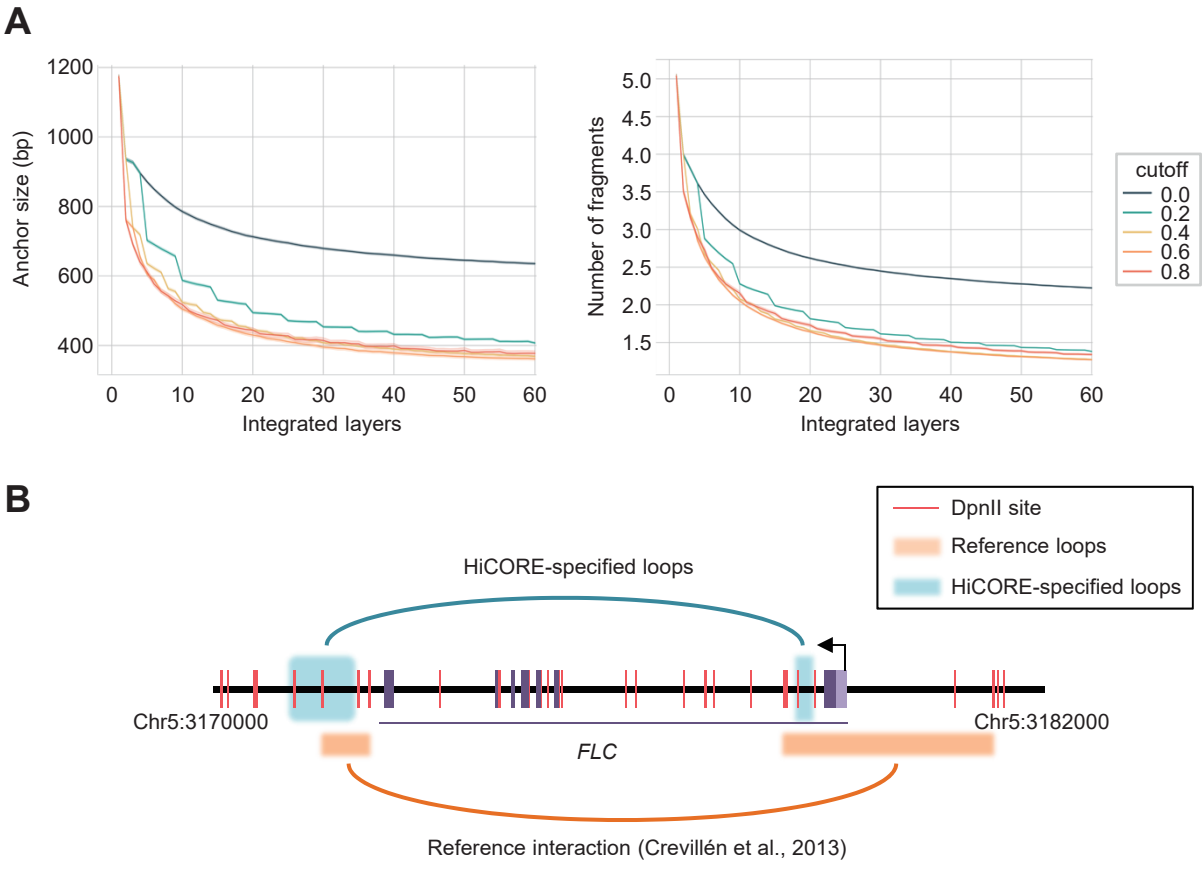

Fig. 7. Application of HiCORE in small genome species *Arabidopsis thaliana*. (A) The average size (bp) and number of fragments in each specified anchor regions. For binning at each layer, restriction fragments smaller than 700 bp were merged with neighboring fragments until the bin is larger than 700 bp. One bin layer was added at a time, repeatedly, and anchor regions were specified. The x-axis indicates the number of layers integrated for HiCORE analysis. Cutoff values indicate the layer-detection frequency. (B) The empirically-proven chromatin loop in the *Arabidopsis FLC* locus <Chr5:3172571-3173171>:<Chr5:3178619-3181368>. Red lines indicate the MboI restriction sites. HiCORE-specified loops were compared with the experimentally validated reference loop.

HiCORE can be incorporated into current Hi-C data analysis pipelines, improving the resolution in predicting chromatin looping regions. HiCORE analysis will facilitate better separation of genomic contacts and the identification of potential genomic elements.

Software availability

HiCORE is implemented as a Python3 package. The most current stable version is available at: https://github.com/CDL-HongwooLee/HiCORE. Detail information to implement HiCORE algorithm is described in (Supplementary Text S1).

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

We thank Dr. Sangrea Shim (Seoul National University, Korea) for fruitful discussion. This work was supported by the Basic Science Research (NRF-2019R1A2C2006915) and Basic Research Laboratory (2020R1A4A2002901) programs provided by the National Research Foundation of Korea and by Creative-Pioneering Researchers Program through Seoul National University (0409-20200281).

AUTHOR CONTRIBUTIONS

P.J.S. and H.L. conceived and designed the analysis. H.L. developed software and performed analysis. P.J.S. wrote the paper with the help of H.L.

CONFLICT OF INTEREST

The authors have no potential conflicts of interest to disclose.

ORCID

Hongwoo Lee https://orcid.org/0000-0002-9339-2757
Pil Joon Seo https://orcid.org/0000-0002-5499-3138

REFERENCES

Cameron, C.J.F., Dostie, J., and Blanchette, M. (2020). HIFI: estimating DNA-DNA interaction frequency from Hi-C data at restriction-fragment resolution. Genome Biol. *21*, 11.

Cao, Y., Chen, Z., Chen, X., Ai, D., Chen, G., McDermott, J., Huang, Y., Guo, X., and Han, J.D.J. (2020). Accurate loop calling for 3D genomic data with cLoops. Bioinformatics *36*, 666-675.

Choi, W.Y., Hwang, J.H., Cho, A.N., Lee, A.J., Jung, I., Cho, S.W., Kim, L.K., and Kim, Y.J. (2020). NEUROD1 intrinsically initiates differentiation of indeuced pluripotent stem cells into neural progenitor cells. Mol. Cells *43*, 1011-1022.

Crevillén, P., Sonmez, C., Wu, Z., and Dean, C. (2013). A gene loop containing the floral repressor FLC is disrupted in the early phase of vernalization. EMBO J. *32*, 140-148.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306-1311.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., 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.

Feng, S., Cokus, S.J., Schubert, V., Zhai, J., Pellegrini, M., and Jacobsen, S.E. (2014). Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in *Arabidopsis*. Mol. Cell *55*, 694-707.

Fernandez-Albert, J., Lipinski, M., Lopez-Cascales, M.T., Rowley, M.J., Martin-Gonzalez, A.M., del Blanco, B., Corces, V.G., and Barco, A. (2019). Immediate and deferred epigenomic signatures of in vivo neuronal activation in mouse hippocampus. Nat. Neurosci. *22*, 1718-1730.

Heidari, N., Phanstiel, D.H., He, C., Grubert, F., Jahanbanian, F., Kasowski, M., Zhang, M.Q., and Snyder, M.P. (2014). Genome-wide map of regulatory interactions in the human genome. Genome Res. *24*, 1905-1917.

Kaul, A., Bhattacharyya, S., and Ay, F. (2020). Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. Nat. Protoc. *15*, 991-1012.

Lajoie, B.R., Dekker, J., and Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods *72*, 65-75.

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.

Liu, C., Wang, C., Wang, G., Becker, C., Zaidem, M., and Weigel, D. (2016). Genome-wide analysis of chromatin packing in *Arabidopsis thaliana* at single-gene resolution. Genome Res. *26*, 1057-1068.

Liu, Q., Lv, H., and Jiang, R. (2019). hicGAN infers super resolution Hi-C data with generative adversarial networks. Bioinformatics *35*, i99-i107.

Liu, T. and Wang, Z. (2019). HiCNN: a very deep convolutional neural network to better enhance the resolution of Hi-C data. Bioinformatics *35*, 4222-4228.

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.

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.

Ong, C.T. and Corces, V.G. (2009). Insulators as mediators of intra- and inter-chromosomal interactions: a common evolutionary theme. J. Biol. *8*, 73.

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.

Roayaei Ardakany, A., Gezer, H.T., Lonardi, S., and Ay, F. (2020). Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. Genome Biol. *21*, 256.

Rowley, M.J., Poulet, A., Nichols, M., Bixler, B., Sanborn, A., Brouhard, E., 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.

Sun, L., Jing, Y., Liu, X., Li, Q., Xue, Z., Cheng, Z., Wang, D., He, H., and Qian, W. (2020). Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in *Arabidopsis*. Nat. Commun. *11*, 1886.

Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.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.

Wang, C., Liu, C., Roqueiro, D., Grimm, D., Schwab, R., Becker, C., Lanz, C., and Weigel, D. (2015). Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*. Genome Res. *25*, 246-256.

Zhang, Y., An, L., Xu, J., Zhang, B., Zheng, W.J., Hu, M., Tang, J., and Yue, F. (2018). Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. Nat. Commun. *9*, 750.