Spatial chromatin accessibility sequencing resolves high-order spatial interactions of epigenomic markers

Yeming Xie[1,†], Fengying Ruan[1,†], Yaning Li[1,†], Meng Luo[1], Chen Zhang[1], Zhichao Chen[1,2], Zhe Xie[2,3], Zhe Weng[1], Weitian Chen[1,2], Wenfang Chen[1], Yitong Fang[1], Yuxin Sun[1], Mei Guo[1], Juan Wang[1], Shouping Xu[4], Hongqi Wang[1,*], Chong Tang[1,*]

[1]BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China
[2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Department of Biology, Cell Biology and Physiology, University of Copenhagen 13, 2100 Copenhagen, Denmark
[4]Department of Breast Surgery, Harbin Medical University Cancer Hospital, Harbin, 150040, China

[†]These authors contributed equally to this work.

Keywords: Pore-C, Hi-C, chromatin accessibility, accessible chromatin, inaccessible chromatin, interaction, methylation, m6A, GpC, CpG

Running title: Multi-omics Pore-C contact map provides chromatin accessibility information

[*]Correspondence:

Chong Tang, email: tangchong@bgi.com; Hongqi Wang, email: hwang@bgi.com

Contact author:

Chong Tang

Director of Technology, BGI-Shenzhen, China

Phone: +8618025420976

Email: tangchong@bgi.com

Abstract

As the genome is organized into a three-dimensional structure in intracellular space, epigenomic information also has a complex spatial arrangement. However, most epigenetic studies describe locations of methylation marks, chromatin accessibility regions, and histone modifications in the horizontal dimension. Proper spatial epigenomic information has rarely been obtained. In this study, we designed spatial chromatin accessibility sequencing (SCA-seq) to resolve the genome conformation by simultaneously capturing the epigenetic information in single-molecular resolution. Using SCA-seq, we simultaneously disclosed spatial interaction of chromatin accessibility (e.g. enhancer-promoter contacts), CpG island methylation, and spatial insulating functions of the CCCTC-binding factor. We demonstrate that SCA-seq paves the way to explore the mechanism of epigenetic interactions and extends our knowledge in 3D packaging of DNA in the nucleus.

Introduction

The linear arrangement of DNA sequences usually gives an illusion of a one-dimensional genome. However, the DNA helix is folded hierarchically into several layers of higher-order structures that undergo complex spatial biological regulation. The link between gene transcription activity and genome structure was established following an observation that active gene expression proceeds in the decondensed euchromatin, and silenced genes are localized in the condensed heterochromatin [1]. Accessibility of chromatin acts as a potent gene expression regulatory mechanism by controlling the access of regulatory factors [1]. Although this model is attractive, it is simplified in that the genome accessibility is considered only in the horizontal dimension [2]. However, the genome is organized into a three-dimensional structure

inside cells, so the accessibility of genome regions also has similar spatial complexity. For example, promoter accessibility could be regulated by contact with enhancers or silencers [3]. Therefore, sophisticated tools are necessary to obtain information about genome accessibility, depending on the genome organization, to resolve the relationship between chromatin activation and genome interactions.

Most of the tools designed to study chromatin accessibility in the linear form are based on the vulnerability of open/decondensed chromatin to treatment with enzymes such as DNase, micrococcal nuclease (MNase), and transposase. In a pioneer study, Crawford lab used DNase-seq to establish the relationship between DNase-hypersensitive regions and open chromatin [4, 5]. MNase-seq is based on a similar concept [6, 7]. Subsequent studies simplified experiments on chromatin accessibility by taking advantage of the ability of mutant transposase to insert sequencing adapters into open chromatin domains [8, 9]. All these methods rely on statistical calculations of chromatin domain accessibility based on the frequency of enzyme-dependent tags on the accessible genome. Chromatin accessibility has been studied at a single-molecule resolution in recent years to provide insights about chromatin heterogeneity *in vivo*. Approaches such as methyltransferase treatment followed by single-molecule long-read sequencing[10], single-molecule adenine methylated oligonucleosome sequencing assay [11], nanopore sequencing of nucleosome occupancy and methylome [12], single-molecule long-read accessible chromatin mapping sequencing [13 14], and Fiber-seq [15] have been developed for this purpose. Generally, decondensed genomes were methylated using methyltransferases and directly sequenced using third-generation sequencing platforms

(Nanopore, PacBio). These advanced methods offered a single-molecule view of the two-dimensional 2–15 kb long chromatin structures.

However, chromatin has a higher-order organization, and the horizontal dimensional map of chromatin accessibility does not fully reflect reality. Chromosome conformation capture (3C) techniques, for example, Hi-C, SPRITE, CHIAPET, and pore-C, have been widely used to map genome-wide chromatin architecture [16, 17, 18, 19, 20]. Most of these techniques do not allow obtaining epigenome information in the process or examining chromatin conformation with multi-omics. Some advanced approaches, such as Trac-looping[21], OCEAN-C[22], and HiCAR[23], can capture open chromatin and chromatin conformation information simultaneously by enrichment of accessible chromatin regions through solubility or transposons. These advanced approaches were originally developed to enrich the accessible chromatin and efficiently observe the interactions of cis-regulatory elements (cREs). The loss of the condensed chromatin regions and methylation marks restricted the possibility of observing the full-scale genome architecture with multi-omics. Therefore, reconstructing DNA organization with multi-omics information could promote further understanding of the interactive regulation of transcription and enable more detailed studies of DNA interactions.

Here, we developed a novel tool, spatial chromatin accessibility sequencing (SCA-seq), based on methylation labeling and proximity ligation. The long-range fragments carrying the chromatin accessibility, CpG methylation, and chromatin conformation information were sequenced using nanopore technology. We mapped chromatin accessibility and CpG methylation to genome spatial contacts. Heterogeneous chromatin accessibility in proximal interactions suggested complex genome regulation in addition to direct contacts between

genome loci. We believe that SCA-seq may facilitate multi-omics studies of genome spatial organization.

Results

Principle of SCA-seq

Recently, there has been an increasing interest in applying methylation labeling and nanopore sequencing for the analysis of chromatin accessibility at a single-molecule level[10-14, 24]. In this study, we have used SCA-seq to study chromatin spatial density to resolve the genome organization and chromatin accessibility simultaneously. (Fig 1a-1) After cell fixation, we used a methyltransferase enzyme (EcoGII or M.CviPI) to artificially mark accessible chromatin regions. (Fig 1a, 2-3) After the chromatin accessibility information was preserved as m6A or GpC methylation marks, we conducted digestion and ligation steps using chromatin conformation capturing protocols, relying on proximity ligation to ligate together multiple linearly distant DNA fragments that are close to each other in the three-dimensional space. (Fig 1a.4) Next, we performed optimized DNA extraction to obtain pure and large DNA fragments. (Fig 1a.5) The DNA fragments that carried chromatin accessibility information, methylation marks, and three-dimensional conformation information were sequenced using the nanopore method and analyzed in our house pipeline from concatemer alignment to single molecule methylation calling (Fig 1a). The conventional next-generation sequencing-based chromatin conformation protocols only captured the interaction between two genomic loci (Fig 1a). Unlike the conventional protocol, the proximity ligation in SCA-seq, not limited to the first-order ligation, can occur multiple times in one concatemer (genome fragments fixed together as a cluster), informing about the high-cardinality of the genome conformation (Fig 1a, b).

Compared with the output of the comparable techniques Trac-loop [21], Hi-CAR [23], and NicE-C [25], which also capture the accessible chromatin conformation, the SCA-seq preserved more multi-omics information, for example, CpG methylation epigenetic marks, chromatin inaccessibility, and high-order chromatin conformation data (Fig 1b).

First, we experimentally determined the feasibility of SCA-seq. In the methylation reactions, the most suitable methyltransferases, EcoGII and M.CviPI, generated the artificial modifications m6A and m5C(GpC), which are rarely present in the native mammalian genomes [26 27]. Our previous research showed that EcoGII effectively labels accessible chromatin owing to the high density of adenine in the genome [14]. However, when EcoGII was used, the high-density labeled m6A modification either blocked or impaired the activity of the majority of restriction enzymes and produced long-digested fragments, leading to lower chromatin conformation resolution (Sfig 1a, b). We also tried digestion with multiple enzymes, as in Hi-C 3.0 [28], and slightly decreased the fragment length. To solve this problem, we selected the m6A-dependent restriction enzyme DpnI that preferentially digests highly methylated DNA containing methylated adenine and leaves blunt ends. However, such m6A-dependent digestion of the highly methylated accessible chromatin was biased, and the blunt ends were not ligated efficiently. We then tried another approach and used M.CviPI that methylates GpCs (m5C) on the accessible chromatin, and these marks occur four times less frequently than adenosine. In the following steps, DpnII and other enzymes (without GC pattern in the recognition sites) efficiently cut DNA molecules with both methylated and unmethylated GpCs (Sfig 1c–e), followed by ligation. It should be noted that the m5C base-calling algorithm has been gradually improved and is now widely used in nanopore sequencing [29]. Considering the unbiased

digestion, M.CviPI might be a better choice in SCA-seq than EcoGII/DpnI. Next, we analyzed the sequencing data and compared them with those obtained using previous technologies.

SCA-seq has the comparable ability to identify chromatin accessibility and native methylation marks

Our work was inspired by the concepts of nanoNOME-seq, SMAC-seq, and Fiber-seq methods [10-14, 24], which use either M.CviPI or EcoGII methyltransferases to label chromatin-accessible regions with methylation sites. Our previous experiments[14] and validations of the results against published data confirmed the effectiveness of the methyltransferase-mediated labeling, showing technological advantages of the complex genome alignment and single-molecule resolution [24]. As the SCA-seq generated discontinuous genomic segments by ligating fragments (Fig 1), which might affect data processing, we first validated the accuracy of methylation calling and methyltransferase labeling in SCA-seq.

First, we performed the initial quality control of the sequencing data for HEK293 cells by validating the methylation calling. We generated 129.94 Gb (36.9× coverage) of mapped sequencing data with an N50 read length of 4,446 bp. To obtain the methylation information from the nanopore data, we adopted well-established methylation caller Nanopolish with the cpggpc calling module[12] and achieve considerable success (AUC CpG = 0.908, GpC = 0.984). In the further validation of methylation calling, we parallelly performed the gold standard whole-genome bisulfite sequencing. The results of the whole-genome bisulfite sequencing analysis were highly correlated with those of Nanopolish (Sfig 2), supporting the accuracy of the methylation caller. In addition to the methylation calling accuracy, there might also be some ambiguity between the native and artificially labeled cytidine methylations. We first checked the native or false-positive GpC regions, which were also very rare and accounted for only 1.8% in

the unlabeled genome (Sfig 3a). GpC ratio of the M.CviPI-treated genomes were significantly higher than the background GpC ratio (Sfig 3b, >10 fold change, $< 2.6\times10^{-16}$, Student's *t*-test). Second, the GCG pattern in the genome might also cause the ambiguity of native methylation CpG or accessibility representing GpC; therefore, we excluded both CpG and GpC methylations from the GCG context (5.6% of GpCs and 24.2% of CpGs) to obtain unbiased methylation information. The excluded GCGs did not significantly affect most of the biological methylation analysis, as reported previously [12]. In conclusion, our bioinformatic pipeline was able to detect native CpG methylations and artificially labeled GpC methylations.

We next assessed the potential of SCA-seq to reveal simultaneously endogenous methylation and chromatin accessibility. As the ATAC-seq and DNase-seq are gold standards for detecting chromatin accessibility, we compared the labeling accuracy of SCA-seq with that of ATAC-seq/DNase-seq globally and locally. Of the 55,429 accessibility peaks called from the SCA-seq data in the whole genome, 74.6% overlapped with peaks observed in ATAC-seq or DNase-seq (Fig 2a). These results were comparable to those of a previous study, in which methyltransferase labeling was used [12]. The ATAC-seq/DNase-seq unique peaks were less frequent, as indicated by the larger *P*-value calculated by MACS2 (Fig 2b). On the other hand, the sequencing depth could improve the sensitivity of SCA-seq to identify the less frequent peaks (Sfig 4de). Previous publications also suggested that the difference between the outputs of the Nanopore-based and next-generation sequencing-based methods might be explained by sequencing depth [12,13]. In the local comparison, SCA-seq also showed peak patterns around the ATAC-seq-identified peaks (Fig 2c). Moreover, we computationally predicted binding sites of the CCCTC-binding factor (CTCF), which usually associates with open chromatin [30]. As

expected, the native CpG methylation level decreased, whereas GpC accessibility increased around the CTCF-binding sites (Fig 2d). At active human transcription start sites (TSSs) with high expression, “open” chromatin regions hypersensitive to transposon attack were observed in ATAC-seq/DNase-seq. SCA-seq showed similar nucleosome depletion patterns around TSSs (Fig 2e). Inactive TSSs (low quantile of gene expression) were less accessible than active TSSs (upper quantile of gene expression) (Fig 2e). In the detailed examination of the genome regions, the SCA-seq showed the expected nucleosome pattern co-localizing with various epigenetic marks, for example, H3K4me3 (active) and H3K27ac (active) (Fig 2f and Sfig 5). To further estimate the labeling efficiency between different methods, we also compared the fold-change values of signal enrichment around TSS with HiCAR, which enriched the accessible chromatin conformation (Sfig 4a). Both SCA-seq and HiCAR showed similar labeling efficiency on the TSS. We then explored the time-dose influence on the labeling results. The relationship between the dose and M.CviPI treatment effect demonstrated superior efficiency of the 3 h treatment, comparing with those achieved after 15 or 30 min treatment (Sfig 4b,c). Overall, SCA-seq reliably estimated chromatin accessibility at the genome level.

SCA-seq reveals high-order chromatin organization

In addition to the methylation information, SCA-seq also preserved genome spatial structure. Therefore, we next validated genome spatial organization. First, we analyzed basic statistical parameters, for example, contact distance and cardinality of SCA-seq. As SCA-seq ligated the multiple fragments together, revealing the multiplex chromatin conformation, we aligned non-singleton chimeric reads into genomic segments and assembled *in silico* paired-end tags

(PETs) in order to compare with Hi-C carrying paired loci. The segment median length was approximately 700 bp (Sfig 6a). Among the informative intra-chromosome PETs, 0.1% of the PETs (contact distance) were <150 bp; 0.3% of them ranged from 150 to 1,000 bp; 24.5% were 1,000–200,000 bp; and 75.1% were >200,000 bp. Unlike Hi-C, SCA-seq, derived from pore-C, revealed the multiplex nature of chromatin interactions. As for the intra-chromosome interactions, 14.7% of the reads contained two segments (cardinality = 2); approximately 14.5% of the reads contained 3–5 segments (cardinality = 3–5); and 5.4% of the reads had more than five segments (cardinality > 5) (Sfig 7a). As in the previous report [31], most of the contacts from the reads with fewer segments appeared to have closer contact distance (Sfig 7b). The contacts from the reads with more segments appeared to have more distal interactions (Sfig 7c, d). The high cardinality of concatemers with enriched enhancer and/or promoter might indicate the cooperativity in the mammalian transcriptional regulation, as in previous reports [31]. Compared with the output of similar methods, such as Trac-loop and HiCAR, SCA-seq also resolved more high-cardinality chromatin conformation and distal interactions (Sfig 7e, f, g). Our results regarding high-order chromatin conformation capturing also agreed with the output of previously published methods, such as SPRITE [18], Pore-C [31], and ChIA-Drop [32], which also disclosed more distal interactions.

We then compared SCA-seq to the gold standard Hi-C with respect to the false positive call rate, reproducibility, and ability to resolve genome spatial organization. False positive call rates of SCA-seq and Hi-C, inferred from hybrid PETs that consisted of mitochondrial DNA and genomic DNA, were similar (Sfig 6b). The compartment score correlation between SCA-seq replicates and pore-C replicates was approximately 0.94, suggesting comparable

reproducibility (Sfig 6d). Furthermore, SCA-seq revealed genome organization similar to the one detected using *in situ* Hi-C. Side-by-side visualization of interaction heatmaps, loops, topologically associating domain (TAD) boundaries, and A/B compartments obtained using SCA-seq and Hi-C showed equivalent genome organizations (Fig 3b, c, e, g, h). The correlation coefficients of the eigenvector and insulation scores were 0.91 and 0.84, i.e., slightly lower than we expected (Fig 3d, f). The sequencing depth of SCA-seq could improve these correlations with Hi-C results (Sfig 6c). Because SCA-seq is a non-amplification method, whereas Hi-C is an amplification method, these differences might have resulted from the amplification bias of GC regions (Sfig 6e, g, f, h). Notably, we found that 66% of the concatemers were compartment-specific (all the segments in one concatemer belonged to either compartment A or B), and 34% were non-specific (mixed composition of segments from compartments A and B) [33]. Overall, these results suggested that SCA-seq successfully resolved the multiplex nature of chromatin interactions.

SCA-seq resolves the relationship between transcription regulator binding and chromatin conformation

We also could observe chromatin conformation with a specific binding pattern from SCA-seq, for example, the CTCF binding pattern. As previous publications mentioned, occupation of binding sites by CTCF could lead to the emergence of short regions (~50 bp) inaccessible to methyltransferase labeling [34, 35], indicating the CTCF binding status on the CTCF motif loci. As expected, SCA-seq also could resolve the transcription factor-specific (~50 bp peak) and nucleosome footprints, as was described previously [34, 35] (Fig 3i). Based on the specific accessibility patterns, we classified chromatin interaction concatemers containing CTCF motifs

into two classes: with and without a CTCF footprint (Fig 3j). Considering the relationship between CTCF binding and chromatin structure formation [36], we plotted the concatemer cardinality and interaction distance (Fig 3k, i). We found that the CTCF binding resulted in higher cardinality and more distal interactions than the non-CTCF binding, suggesting that CTCF binding facilitates formation of a more complex structure. As has been reported recently [35], the methyltransferase accessibility pattern also could reflect other transcription factors' footprints, enlightening the further exploring the relationship between chromatin conformation with other transcription factors by SCA-seq. Therefore, SCA-seq could help to subgroup chromatin interaction concatemers and investigate the relationship between chromatin conformation and protein binding.

SCA-seq resolves spatial interactions of accessible and inaccessible chromatin regions

Given the genome organization is highly heterogenous in different cells, our chromatin interaction status analysis mainly relied on the single-molecule pattern, which needs high sensitivity and specificity. Single-molecule base modification calling was performed as described previously [12]. Moreover, we determined enzyme labeling efficiency, which was 79–88%, based on the CTCF motifs and the Lambda DNA spike-in control measurements (Sfig 3b,c). Next, we filtered the segments using the binomial test to minimize false positive attribution of the accessible or inaccessible status (see Methods). Unexpectedly, accessible and inaccessible DNAs were ligated together in SCA-seq (Fig 4a), suggesting heterogeneous accessibility of the spatially neighboring DNA regions. Each segment in the concatemer was determined to be either accessible or inaccessible. Then, the fraction of accessible segments in each concatemer was calculated as the $N_{accessible}/(N_{accessible+inaccessible})$. Compartment A had a

significantly higher fraction of accessible segments than compartment B (Sfig 8a, b). Our overall genome concatemer calculations showed that 29% of the genome concatemers were inaccessible on all enclosed segments (the fraction of accessible segments < 0.1). Furthermore, 62.2% of genome concatemers had both accessible and inaccessible segments (hybrid concatemers), and only 8.8% maintained all segments as accessible (the fraction of accessible segments > 0.9) (Fig 4b). Figure 4a illustrates an example region with 1D genome feature tracks to demonstrate the promoter-enhancer spatial interactions, accessibility, and CpG methylation at a single molecule resolution. nanoNOME-seq that labels chromatin accessibility in single molecules also confirmed the existence of partial hybrid concatemers (Sfig 3d). To explore if the concatemer accessibility status was related to spatial location, we plotted the inaccessible concatemers and hybrid concatemers on the 2D contact heat map. We found that hybrid concatemers tended to accumulate around the TAD boundaries and contain more distant connections (Fig 4b). By plotting the ratio of accessible segments against concatemer cardinality or interaction distance, we revealed that the more accessible segments tended to cluster as high cardinality, also implying their distal and high-cardinality interaction preference on the hybrid concatemers (Fig 4c and Sfig 8e). Moreover, we found that the hybrid concatemers comprised more enhancer and promoter elements than inaccessible concatemers (Fig 4d), suggesting the relationship between concatemer type and the extent of transcription regulation. We further investigated the enhancer and promoter contacts on chromosome 7, 30.3% of which had accessible–accessible status; 18.5% had accessible–inaccessible status, and 51.2% had inaccessible–inaccessible status (Fig 4e). The frequency of contacts with accessible enhancer/promoter was highly correlated with gene expression levels (Fig 4f, g and Sfig 8c, d), supporting the transcription model in which active enhancers

initiate promoters by spatial contact [37]. However, 51.2% of enhancer–promoter interactions were independent of the chromatin-accessible status, suggesting that such spatial interactions were not the only factor regulating the initiation of transcription.

Owing to the close relationship between spatial accessibility and transcriptional activity, we developed an equation to calculate the activation power of genome loci (Sfig 9a, see Methods), revealing their potential to activate transcription. We used SCA-seq to establish the genome activation table between genome bins. The activation matrix times the expected accessibility was equal to the observed chromatin accessibility. Then, we resolved the matrix equation. The expected chromatin accessibility represented the original chromatin accessibility without genome interaction effects. The high signal in expected chromatin accessibility also indicated that this site could potentially activate the contacting sites, which we recognized as the “power center” or “original driver”. Having analyzed the observed chromatin accessibility (Sfig 9b, c), we selected a representative CTCF motif site that activated over 100 contacted sites (Sfig 9d). The locus with higher activation power might enhance genome accessibility of a greater number of contacting loci. Overall, spatial contacts and contact accessibility might cooperatively regulate gene expression levels.

SCA-seq resolves CpG methylation on the spatial contacts with orphan CpG islands

CpG islands (CGI), which are widespread features of vertebrate genomes, were associated with ~50% of gene promoters (pCGI). pCGIs control gene transcription by affecting the neighboring promoters with methylation-triggered chromatin changes. Some CGIs are located close to enhancers. In addition, thousands of orphan CGIs (oCGIs), which are at a longer

distance (1 kb) from the promoters and enhancers, have been barely unknown (Fig 5a) [38, 39]. In the SCA-seq data that indicated the high-order interaction and CpG methylation, we found 76,418 reads overlapping with CGIs on chromosome 7, and the majority of oCGIs were usually spatially close to the CTCF binding motifs and active histone markers, such as H3K27ac and H3K4me3, suggesting their active regulatory functions (Fig 5b). By examining the methylation status on reads, as expected, these read segments demonstrated lower CpG methylation and higher chromatin accessibility (GpC methylation), which further supports their roles in gene activation (Fig 5b). In a previously published study[38], oCGIs were considered to act as tethering elements that promote topological interaction between enhancers and distally located genes to regulate gene expression. In SCA-seq, we observed that 60% of oCGIs tethered at least one type of regulatory elements, such as enhancers, CTCFs, and promoters (Fig 5c). After normalizing by the total number of regulatory elements, we found that the oCGIs preferentially interacted with CTCF and promoters, comparing with non-CGIs ($P < 2.6 \times 10^{-16}$, binomial test, background frequency was used as control) (Fig 5d). Further analysis of each concatemer type showed that 39% of oCGI-enhancer concatemers and oCGI-CTCF concatemers included more than two enhancers or CTCF motifs. In contrast, most of the oCGIs tethered to one promoter (Fig 5e). However, we found that CpG methylation on oCGIs was weakly correlated with the chromatin accessibility of promoters in regression analysis, whose mechanism of regulation need to be further studied. Overall, oCGIs were found to tether the enhancers and CTCF motifs to communicate with the promoters, which extends our understanding of oCGI regulatory functions.

Discussion

We developed SCA-seq to expand the notion of traditional chromatin accessibility to high dimensional space by simultaneously resolving chromatin accessibility and genome conformation. Compared with 1D ATAC-seq, SCA-seq might more closely represent the true structure of the native genome. With SCA-seq, we found that genome spatial contacts maintained non-uniform chromatin accessibility, suggesting complex genome regulation in the 3D space. SCA-seq could be a multi-omics tool to simultaneously analyzed the DNA methylation, chromatin accessibility, transcription factor binding and chromatin conformation. For the single-molecule resolution, the efficiency of methyltransferase labeling must be ensured. We used lambda DNA *in vitro* labeling and *in vivo* CTCF motif signal to estimate labeling efficiency. The binomial test was utilized to correct the labeling accuracy at a single-molecule level. Then, relatively reliably accessible chromatin markers were obtained. However, it is still possible for such a marker to be missed or overridden in the case of insufficient enzyme activity. Because suboptimal labeling efficiency may lead to deviations from our conclusions, our analysis in the following experiments was mainly based on the statistics of large numbers of molecules. In the single-molecule analysis of specific locations, more than two similar concatemers could accurately describe the epigenetic status in the exact spatial locations. Given the high heterogeneity of the dynamic genome structure and SCA-seq resolution, a much higher sequencing throughput is required to achieve analysis at a single-molecule level in a specific spatial location.

The eigenvalue and insulation score of SCA-seq moderately correlated with those of gold standard Hi-C (0.91 and 0.84), and the moderate correlations were also true in all other multiplex-order chromatin conformation methods, such as Pore-C [20], SPRITE [18], and ChIA-

Drop [32]. After we studied this problem deeper, we found that there may be two reasons for this. First, the conversion from multi-contacts to pairwise contacts overrepresented some interactions. In our algorithm, we significantly resolved this issue by weighted transformation, which increased the correlation coefficient to 0.9. Second, we found that the low correlation regions had lower GC density and low read counts (Sfig 5e, f, g, h). It has been pointed out previously[40] that PCR amplification could bias eigenvalues and insulation scores. In contrast, Pore-C and SCA-seq are non-amplification methods. After the quality filter of low-coverage regions, the correlation between the two methods was significantly improved.

SCA-seq was created as a multi-omics tool to examine both chromosome conformation and chromatin accessibility. It is important to discuss different levels of resolution of chromosome conformation capture and chromatin accessibility. The resolution of the chromosome conformation capture is approximately 700 bp, whereas that of the conventional chromatin accessibility is approximately 200 bp. Not all precise accessible–accessible chromatin interactions can be determined. Therefore, an alternative hypothesis is that the interaction loci are located outside the accessible chromatin. Therefore, improvement of the resolution of chromosome conformation capture in SCA-seq is needed to determine spatial accessibility interaction accurately.

Overall, our results demonstrated that SCA-seq could resolve genome accessibility locations in the three-dimensional space, facilitating the observation of subgroups of chromatin conformation regions with a specific binding pattern, conformation-based chromatin accessibility, and conformation-based native CpG methylation. SCA-seq might pave the way to explore dynamic genome structures in greater detail.

Method

The detailed protocol could be found https://www.protocols.io/view/sca-seq-b6a6rahe. The bioinformatic script could be found https://github.com/genometube/SCA-seq. The data source and QC information could be found in the supplemental files.

Cell culture

Derivative human cell line which expresses a mutant version of the SV40 large T antigen (HEK 293T) [abclonal] was maintained in DMEM-high glucose [Thermo Fisher 11995065] supplemented with 10% fetal bovine serum (FBS) [Thermo fisher 1009141].

Cross-linking

5 million cells were washed 1 time in chilled 1X phosphate buffered saline (PBS) in a 15 mL centrifuge tube, pelleted by centrifugation at 500xg for 3 min at 4℃. Cells were resuspended by gently pipetting in 5 mL 1X PBS with formaladehyde (1% final concentration). Incubating cells at room temperature for 10 min, add 265 µL of 2.5 M glycine (125 mM final concentration) and incubate at room temperature for 5 min to quench the cross-linking. Centrifugate the mix at 500xg for 3 min at 4℃. Wash cells 2 times with chilled 1X PBS.

Nuclei isolation and methylation

Cell pellet was resuspended with cold lysis buffer: 10 mM HEPES-NaOH pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 1X proteinase inhibitor [Sigma 11873580001], 0.1% Tween-20, 0.1 mg/ml BSA, 0.1 mM EDTA, 0.5% CA-630, incubate on ice for 5 min. Centrifugate lysis mixture at 500xg for 5 min at 4℃ to collect the nuclei. Washed the nuclei once with 1X GC buffer [NEB M0227L] then resuspend 2 million nuclei in 500 µL methylation reaction mixture: 1X GC buffer, 200 U M. CvipI [NEB M0227L], 96 µM S-adenosylmethionine, 300 mM Sucrose, 0.1 mg BSA, 1X proteinase inhibitor, 0.1% Tween-20. Incubate the reaction for 3 hours at 37℃, add 96 µM SAM and 20 U M.CvipI per hour. Centrifugate at 500xg for 10 min at 4℃ to collect nuclei, wash the nuclei once with chilled HEPES-NaOH pH7.5 and centrifugate to collect nuclei.

Restriction enzyme digest

Resuspend nuclei with 81 µL cold HEPES-NaOH pH7.5, add 9 µL 1% SDS and react at 65℃ for 10 min to denature the chromatin, take the tube on ice immediately after the reaction. We also tried 0.5% SDS, and the results were similar (Data not shown). Add 5 µL 20% Triton X-100 and incubate on ice for 10 min to quench SDS. Prepare digestion mixture: 140 U DpnII [NEB R0543L], 14 µL 10X HEPES-buffer3.1 [50 mM HEPES-NaOH pH 8.0, 100 mM NaCl, 10 mM $MgCl_2$, 100 µg/mL BSA], add nuclei suspension and nuclease-free water into mixture to achieve a final volume of 140 µL. Incubate digest mixture in a thermomixer at 37℃ for 18 hours with 900 rpm rotation.

Ligation

DpnII digests were heat inactivated at 65℃ for 20 min with 700 rpm rotation, average digests to 70 μL per tube, add 14 μL T4 DNA Ligase buffer [NEB M0202L], 14 μL T4 DNA Ligase [NEB M0202L], 1 mM ATP and nuclease-free water to achieve a final volume of 140 μL. The ligation was incubated at 16℃ for 10 hours with 800 rpm rotation.

Reverse cross-linking and DNA purification

Collect all ligation into one 1.5 mL tube, add equal volume of 2X sera-lysis [2% Polyvinylpyrrolidone 40, 2% Sodium metabisulfite, 1.0 M Sodium Chloride, 0.2 M Tris-HCl pH 8.0, 0.1 M EDTA, 2.5% SDS], add 5 μL RNaseA [QIAGEN 19101], incubate at 56℃ for 30 min. Add 10 μL Proteinase K [QIAGEN 19131], 50℃ overnight incubation with 900 rpm rotation. DNA was purified with high molecular weight gDNA extraction protocol [Baptiste Mayjonade, 2016].

Nanopore sequencing

The DNA from the reactions were purified, and library were prepared following the manufacturer's protocol of SQK-LSK109 (Nanopore, SQK-LSK109). The library was sequenced in the ONT PromethION platform with R9.4.1 flow cell.

WGBS

The WGBS was parallelly performed from the purified DNA product by using MGIEasy Whole Genome Bisulfite Sequencing Library Prep Kit (MGI 1000005251). The final products were sequenced by MGISEQ-2000.

SCA-seq pipeline.

We developed a reproducible bioinformatics pipeline to analyze the M.CvipI footprint and CpG signal on SCA-seq concatemers. Briefly, the workflow starts with the alignment of SCA-seq reads to a reference genome by bwa (v0.7.12) using the parameter bwa bwasw -b 5 -q 2 -r 1 -T 15 -z 10. The mapping score ≥ 30, and reads with length <50 bp were set to filter out the low-quality mapping segment. To remove the non-chimeric pairs due to ligation of cognate free ends or incomplete digestion, each alignment is assigned to an in-silico restriction digest based on the midpoint of alignment. The locus of each segment on each concatemer is summarize by converting the filtered alignment to a segment bed file sorted by read ID first and then the genome locus. The alignment bam file is also used to call the GpC and CpG methylation by Nanopolish (v0.11.1) call-methylation with the cpggpc model (--methylation cpggpc). The default cut-off for log-likelihood ratios are used to determine methylated GpC (>1) and methylated CpG sites (>1.5) [12]. The methylation call is then counted to each segment in the segment bed file to derive the methylated and unmethylated count of GpC and CpG for each segment of the concatemers.

SCA-seq and Hi-C comparisons

SCA-seq concatemers were converted into virtual pairwise contacts in order to correlate with the published Hi-C datasets. For a given pair of two segments in a SCA-seq concatemer, the number of segments between the two segments (k) is positively correlated with the spatial distance between them and negatively correlated with the Hi-C observable counts. Therefore, we down-weighted each of the two segments in a SCA-seq concatemer such that each pair of two segments has a weight of $1/2^k$. The decomposed SCA-seq contact matrix was treated as a Hi-C contact matrix and analyzed by Cooltools (v0.4.1) [41].The contact matrix was normalized using cooler balance. Then the eigenvector scores and TAD insulation score were calculated by Cooltools call-compartments and Cooltools diamond-insulation tools. The linear correlation between the Pore-C and Hi-C contact matrices was then measured by eigenvector scores (compartment score) and TAD insulation score calculated by Cooltools. The variation of individual pore-C runs, individual SCA-seq runs, and downsampled SCA-seq datasets were also examined by the above metrics. Loop anchors were identified by ENCODE CTCF ChiP-seq peaks (ENCSR135CRI). Cooltools pileup was used to compute aggregate contact maps at 10kb resolution and centered at the loop anchors (± 100kb).

SCA-seq, ATAC-seq, and DNase-seq comparisons

For comparison and visualization of bulk accessibility, the conventional bulk ATAC-seq and DNase-seq data of HEK293T peak signals were obtained from Gene Expression Omnibus (GEO) accession GSE108513 and GSM1008573. The SCA-seq accessibility peak calling was performed in a similar way to nanoNOMe [12]. Briefly, 200bp window and 20bp step size continuous regions of GpC methylated counts, unmethylated counts, and GpC methylation

ratio were generated from SCA-seq Nanopolish calls. The regions of GpC methylation ratio greater than 99th percentile of the regions were selected as candidate first. The significance of each candidate region was calculated by the one-tailed binomial test of raw frequency of accessibility (methylated GpC site / total GpC site) to reject the null probability, which is defined by the overall regions GpC methylation ratio. The p-values were corrected for multiple testing by Benjamini-Hochberg correction. The adjusted p-values < 0.001 and widths greater than 50 bps were determined as the SCA-seq accessibility peaks. The overlapping peaks between SCA-seq, ATAC-seq, and DNase-seq were identified by bedtools (v2.26.0) intersect.

Analysis of cytosine modifications called by WGBS

The cytosine modification analysis of WGBS data generally follows the rules in Heyn et al 2012 [42]. In brief, the two sets of hg19 genome reference sequences were prepared; the C to T reference that had the C residues replaced by Ts, and the G to A reference that had the T residues replaced by As. The two sets of reads are prepared the same way; the C to T reads and G to A reads. The two sets of reads were aligned to the two sets of genome separately by GEM alignment software [43] which allows 4 mismatches of bases with quality score over 25. The lack of methylation of C residues would be recognized as C to T or G to A conversions. The methylation ratio of methylated GpC or CpG residues in 200bp windows were then calculated for the correlation analysis with the SCA-seq methylation calling.

Estimate the labeling efficiency *in vivo*

As previous research, the CTCF motif maintained the accessible chromatin in neighboring 200bp region. Consider the resolution in Hi-C and experimental fragmentation, we selected the 1000bp bins with the documented CTCF motif in center. The CpG methylation levels were negatively correlated with the chromatin accessibility. Then the segment with low CpG methylation were expected to maintain the accessible chromatin status with CTCF binding. We hypothesized that the segments with low CpG methylation (CpG ratio<0.25) and low chromatin accessibility (GpC ratio<0.1) were not efficiently labeled.

Filter the segments by binomial test

The medium segment length is 500bp, which is close to the general size of accessible chromatin segments. We first calculated the background level of the methyl-GpC (accessible) and non-methyl-GpC (inaccessible) probability on the segments. We used the non-treated genomic DNAs as the background, and 0.03 (GpC background) were the average GpC frequency on segments. Then we performed the binomial test (R basics) for each segments in M.CviPI treated samples to test the null hypothesis that if labeled GpCs (GpC>=4) was equal or smaller than the background GpCs. We further to investigate the confidence level of inaccessible chromatin with the non-methyl GpC. The non-methyl-GpC frequency in M.CviPI treated spike-in is 0.3. Therefore, we roughly estimated that 21% GpCs(p) were not efficiently labeled by M.CviPI. Then we performed the binomial test (R basics) for each segment in M.CviPI treated samples to test the null hypothesis that if the non-methyl GpCs on heterochromatins were equal or larger than the enzymatic inefficiency. For both p-value, the probabilities were corrected for multiple testing using the Benjamini Hochberg correction and

accessible/inaccessible segments with adjusted p-value less than 0.05. We determined the accessible segments first, and then we further determined the inaccessible segments in the rest. There are ~2 millions segments which is undetermined and discarded.

High resolution accessibility determination

As above description, we used the binomial test to test the accessibility on each segment. However, the accessible regions in ATAC-seq were around 200bp (peak average size). If we used sliding windows (200bp windows, sliding 50bp) on each segment, we may determine the precise accessible regions on the segments with sacrificing the computational speed. By the similar methods, we performed the binomial test (R basics) for each windows in M.CviPI treated samples to test the null hypothesis that if labeled GpCs (GpC_methy>1) was equal or smaller than the background GpCs. We defined the accessible segments as containing >=1 accessible windows. Finally, we found that the sliding windows methods could produce 8% more accessible, which is not very significant improvement. Considering the general computational ability, we suggested the above methods in our experiments.

Activation power calculation

We used the SCA-seq to establish the genome interaction table between genome bins. In the genome interaction table, the accessible-accessible contacts were defined as 2, which means the contacts could initiate the activation. The inaccessible-inaccessible contacts were defined as -2, and the accessible-inaccessible contacts were defined as 0. Then these values were

summarized in the genome interaction table as means. The 3D genome interaction matrix times the expected accessibility was equal to the observed chromatin accessibility (Sfig 12). Then we resolved the matrix equation. The expected chromatin accessibility represented the original chromatin accessibility without the genome interaction effects. The high signal in expected chromatin accessibility also indicated that this site could potentially activate the contacted sites, which we recognized as the “power center.”

Statistics

Most of the parametric data which were distributed as normal distribution (log normal distribution), were performed in two-side *t-test*. The Pearson correlation analysis was also performed for normal distribution data. We used the *Fisher’s exact test* for the differential accessibility analysis in SCA-seq. Other non-parametric or abnormally distributed data were performed as *Wilcoxon rank test*.

Data availability

The data were stored at https://db.cngb.org/search/project/CNP0002862/ and NCBI BioProject PRJNA917827. All custom codes for SCA-seq are available from GitHub https://github.com/genometube/SCA-seq.

Acknowledgment

This research was supported by the Science, Technology, and Innovation Commission of Shenzhen Municipality (grant number JSGG20170824152728492). The supporter had no role in designing the study, data collection, analysis and interpretation, or in writing the manuscript.

Author contributions

CT designed and supervised the experiments. YL, FR, and ML perform the lab experiments; YX and CT perform the bioinformatics data analysis. All authors combinedly performed the data analysis. All authors have read and approved the final manuscript draft.

Competing interest

The authors declare no competing interests.

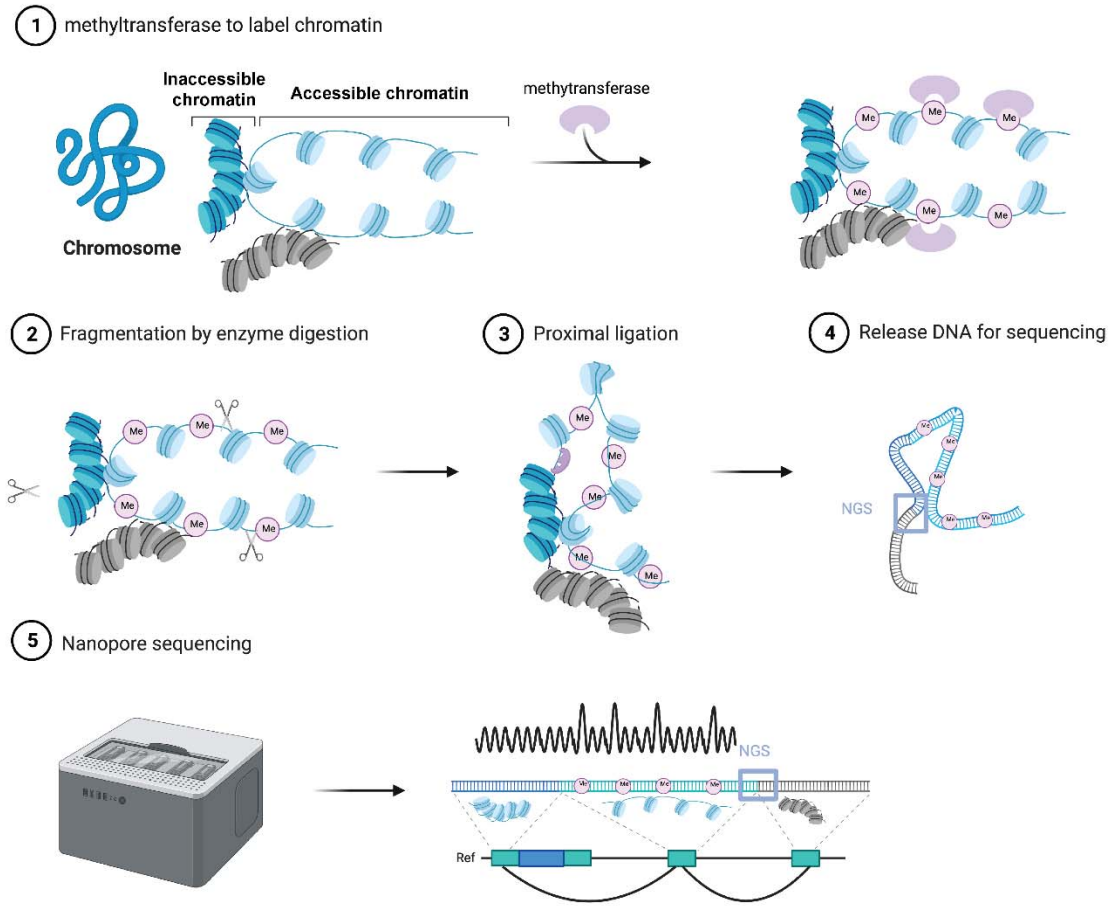
References

1. Klemm, S.L., Shipony, Z. & Greenleaf, W.J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* **20**, 207-220 (2019).
2. Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell* **128**, 787-800 (2007).
3. Kolovos, P., Knoch, T.A., Grosveld, F.G., Cook, P.R. & Papantonis, A. Enhancers and silencers: an integrated and simple model for their function. *Epigenetics & Chromatin* **5**, 1 (2012).
4. Boyle, A.P. et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-322 (2008).
5. Song, L. & Crawford, G.E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor protocols* **2010**, pdb.prot5384-pdb.prot5384 (2010).
6. Schones, D.E. et al. Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell* **132**, 887-898 (2008).
7. Henikoff, J.G., Belsky, J.A., Krassovsky, K., MacAlpine, D.M. & Henikoff, S. Epigenome characterization at single base-pair resolution. *Proceedings of the National Academy of Sciences* **108**, 18318-18323 (2011).

8. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**, 1213-1218 (2013).
9. Buenrostro, J.D., Wu, B., Chang, H.Y. & Greenleaf, W.J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**, 21.29.21-21.29.29 (2015).
10. Wang, Y. et al. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Research* (2019).
11. Abdulhay, N.J. et al. Massively multiplex single-molecule oligonucleosome footprinting. *Elife* **9** (2020).
12. Lee, I. et al. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nature Methods* **17**, 1191-1199 (2020).
13. Shipony, Z. et al. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat Methods* **17**, 319-327 (2020).
14. Chen, W. et al. Sequencing of methylase-accessible regions in integral circular extrachromosomal DNA reveals differences in chromatin structure. *Epigenetics & Chromatin* **14**, 40 (2021).
15. Stergachis, A.B., Debo, B.M., Haugen, E., Churchman, L.S. & Stamatoyannopoulos, J.A. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**, 1449-1454 (2020).
16. Bonev, B. et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572.e524 (2017).
17. Rao, S.S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).
18. Quinodoz, S.A. et al. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744-757.e724 (2018).
19. Fullwood, M.J. et al. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64 (2009).
20. Deshpande, A.S. et al. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. *Nature Biotechnology* (2022).
21. Lai, B. et al. Trac-looping measures genome structure and chromatin accessibility. *Nature Methods* **15**, 741-747 (2018).
22. Li, T., Jia, L., Cao, Y., Chen, Q. & Li, C. OCEAN-C: mapping hubs of open chromatin interactions across the genome reveals gene regulatory networks. *Genome Biology* **19**, 54 (2018).
23. Wei, X. et al. Multi-omics analysis of chromatin accessibility and interactions with transcriptome by HiCAR. *bioRxiv*, 2020.2011.2002.366062 (2020).
24. Weng, Z. et al. Long-range single-molecule mapping of chromatin modification in eukaryotes. *bioRxiv*, 2021.2007.2008.451578 (2021).
25. Luo, Z. et al. NicE-C efficiently reveals open chromatin-associated chromosome interactions at high resolution. *Genome Research* (2022).
26. McClelland, M. & Ivarie, R. Asymmetrical distribution of CpG in an 'average' mammalian gene. *Nucleic acids research* **10**, 7865-7877 (1982).
27. O'Brown, Z.K. et al. Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC Genomics* **20**, 445 (2019).
28. Lafontaine, D.L., Yang, L., Dekker, J. & Gibcus, J.H. Hi-C 3.0: Improved Protocol for Genome-Wide Chromosome Conformation Capture. *Curr Protoc* **1**, e198 (2021).

29. Liu, Y. et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome biology* **22**, 295-295 (2021).
30. Ong, C.-T. & Corces, V.G. CTCF: an architectural protein bridging genome topology and function. *Nature Reviews Genetics* **15**, 234-246 (2014).
31. Ulahannan, N. et al. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. *bioRxiv*, 833590 (2019).
32. Zheng, M. et al. Multiplex chromatin interactions with single-molecule precision. *Nature* **566**, 558-562 (2019).
33. Tavares-Cadete, F., Norouzi, D., Dekker, B., Liu, Y. & Dekker, J. Multi-contact 3C reveals that the human genome during interphase is largely not entangled. *Nature Structural & Molecular Biology* **27**, 1105-1114 (2020).
34. Stergachis, A.B., Debo, B.M., Haugen, E., Churchman, L.S. & Stamatoyannopoulos, J.A. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**, 1449-1454 %@ 0036-8075 (2020).
35. Battaglia, S. et al. Long-range phasing of dynamic, tissue-specific and allele-specific regulatory elements. *Nature Genetics* **54**, 1504-1513 (2022).
36. Hyle, J. et al. Acute depletion of CTCF directly affects MYC regulation through loss of enhancer-promoter looping. *Nucleic Acids Res* **47**, 6699-6713 (2019).
37. Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. *Nat Rev Genet* **20**, 437-455 (2019).
38. Pachano, T. et al. Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness. *Nature Genetics* **53**, 1036-1049 (2021).
39. Bell, J.S.K. & Vertino, P.M. Orphan CpG islands define a novel class of highly active enhancers. *Epigenetics* **12**, 449-464 (2017).
40. Niu, L. et al. Amplification-free library preparation with SAFE Hi-C uses ligation products for deep sequencing to improve traditional Hi-C analysis. *Communications Biology* **2**, 267 (2019).
41. Abdennur, N. et al. Cooltools: enabling high-resolution Hi-C analysis in Python. *bioRxiv*, 2022.2010.2031.514564 (2022).
42. Heyn, H. et al. Distinct DNA methylomes of newborns and centenarians. *Proceedings of the National Academy of Sciences* **109**, 10522-10527 (2012).
43. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods* **9**, 1185-1188 (2012).

a



b

	SCA-seq	Trac-loop [11]	Hi-CAR [13]	nanoNome/SMAC/Fiber-seq [5-10]	ATAC-seq [4]	Hi-C [14]
Chromatin conformation signal	Multiple order concatemer (higher-dimension)	Pairewise contact (2D)	Pairewise contact (2D)	Not available (1D)	Not available (1D)	Pairewise contact (1D)
Chromatin accessibility	Direct sequencing of accessible and inaccessible regions	PCR enrichment of accessible regions only	PCR enrichment of accessible regions only	Direct sequencing of accessible and inaccessible regions	PCR enrichment of accessible regions only	Not available
Single molecule resolution	Each read is a multiple order concatemer with accessibility signal for each fragment	Not available	Not available	Each read is a singleton with accessibility signal	Not available	Not available
Read length	Third-generation sequencing to cover full length concatemer	NGS-based short read to partially span the ligation site	NGS-based short read to partially span the ligation site	Third-generation sequencing	NGS-based short read to pile-up on genome locus and call peaks	NGS-based short read to partially span the ligation site
CpG methylation calling	Simultaneous detection of spatial CpG regulation in the same concatemer	Not available	Not available	Simultaneous measurement of CpG in the same singleton molecule	Not available	Not available

Fig 1. Principle of SCA-seq. **(a)** (1) After fixation, chromatin accessibility can be inferred from the extent of artificial methylation by methyltransferase (m6A or GpC, rare in native genomes). (2–4) Restriction enzymes to digest the labeled genome are selected. The interacted segments (>2) stay as a cluster of fragments (**concatemer**) due to fixation. Next, the spatially interacted segments in one

concatemer are proximally ligated, along with the formation of chimeric long fragments. (5) These long chimeric fragments are sequenced using the Nanopore technology. The data-trained Nanopolish model helps calling the labeled artificial methylation marks (GpC or m6A), which reflect the regional chromatin accessibility and native CpG on the chimeric reads. Our algorithm analyzes the composition of the chimeric reads, indicating genome locations of these composed segments. Moreover, SCA-seq can simultaneously capture chromatin accessibility and native methylation information of these segments. **(b)** Comparison of advantages of SCA-seq with those of other similar technologies.

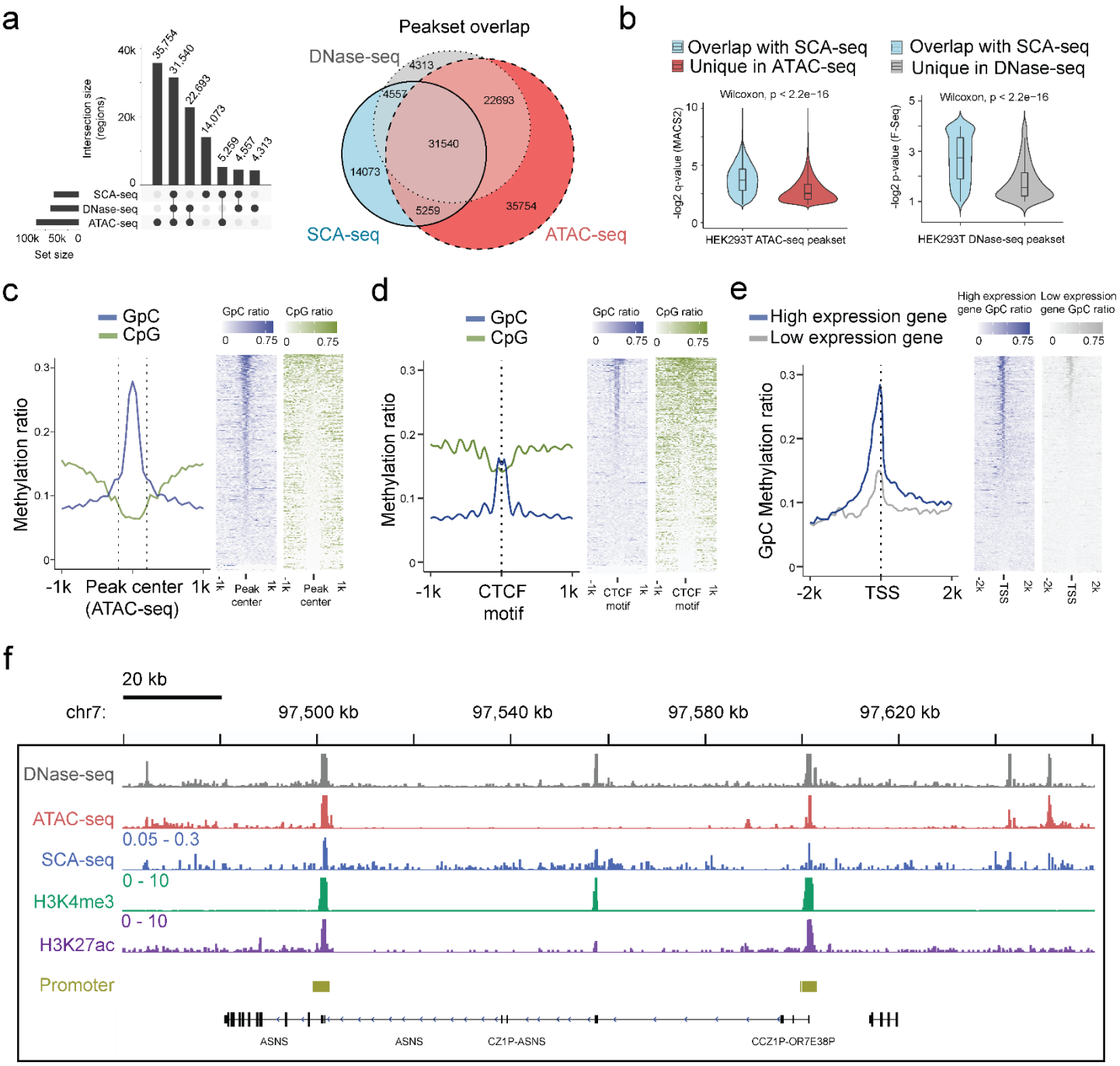


Fig 2. Feasibility of accessible chromatin labeling using SCA-seq. **(a)** Venn diagram of the peak overlap between SCA-seq, ATAC-seq, and DNase-seq. The peak-calling algorithms were borrowed from nanoNOME-seq. There were 72% peaks identified in SCA-seq, overlapping with those detected by either ATAC-seq or DNase-seq. The upset plot is shown on the left. **(b)** Overlapping peaks in the violin plot between SCA-seq and ATAC-seq show the higher MACS2 peak q-value than the non-overlapping peaks ($P < 2.6 \times 10^{-16}$, Wilcoxon rank sum test). Similarly, the overlapping peaks in the violin plot between SCA-seq and DNase-seq have the higher F-seq *P*-value than the non-overlapping

peaks ($P < 2.6 \times 10^{-16}$, Wilcoxon rank sum test). **(c)** ATAC-seq peak centered plot. ATAC-seq peak regions were centered at 0 and are indicated by the dashed line. The Y-axis indicates the average methylation ratio on each site (methylation ratio = methylated GpC/all GpC in a 50 bp bin). Blue and green lines indicate the GpC methylation ratio (accessibility) and CpG methylation ratio, respectively. GpC and CpG ratio heatmaps around ATAC-seq peak centers are shown on the right. Each row represents a region of genome with an ATAC-seq peak. **(d)** CTCF motif centered plot with "0" indicating the CTCF motif position. The Y-axis indicates the average methylation ratio among all the molecules across CTCF motifs. SCA-seq demonstrates classic nucleosome depletion patterns around the CTCF motif. GpC and CpG ratio heatmaps around CTCF sites are shown on the right. Each row represents a region of genome with an CTCF site. **(e)** TSS-centered plot. We classified genes with high expression (upper quantile) and low expression (lower quantile) by the expression ranks. GpC and CpG ratio heatmaps around TSSs are shown on the right. Each row represents a region of genome with a TSS. **(f)** A representative genome browser view showing chromatin accessibility signal of DNase-seq, ATAC-seq, SCA-seq, as well as H3K4me3 and H3K27ac ChIP-seq signals in a genome browser (191 kb span). The SCA-seq track shows the GpC methylation ratio in each genomic locus.

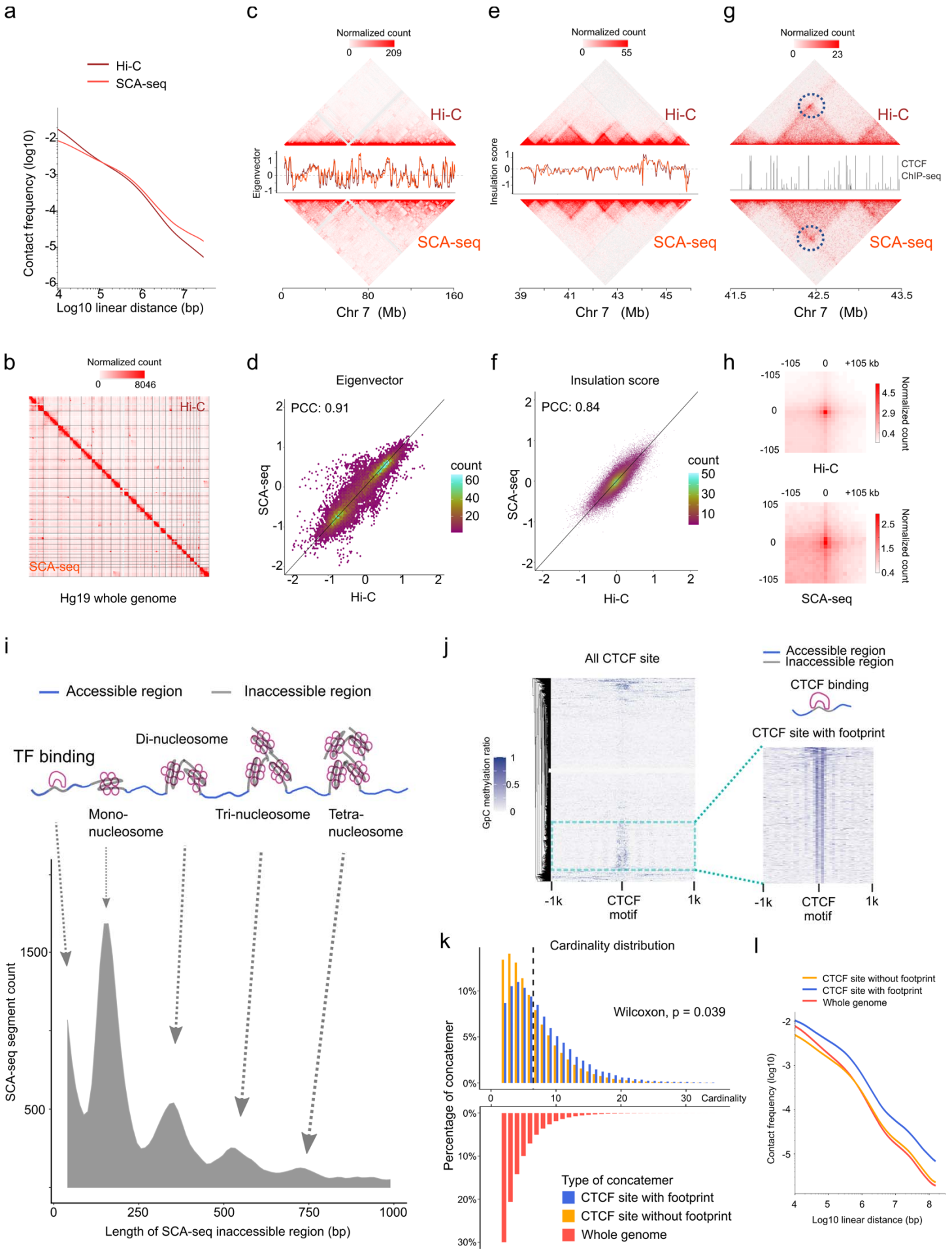

a
Hi-C
SCA-seq
Contact frequency (log10)
Log10 linear distance (bp)
c
Normalized count
0
209
Hi-C
Eigenvector
SCA-seq
Chr 7 (Mb)
e
Normalized count
0
55
Hi-C
Insulation score
SCA-seq
Chr 7 (Mb)
g
Normalized count
0
23
Hi-C
CTCF ChIP-seq
SCA-seq
Chr 7 (Mb)
b
Normalized count
0
8046
Hi-C
SCA-seq
Hg19 whole genome
d
Eigenvector
PCC: 0.91
SCA-seq
Hi-C
count
f
Insulation score
PCC: 0.84
SCA-seq
Hi-C
count
h
-105
0
+105 kb
Normalized count
Hi-C
SCA-seq
i
Accessible region
Inaccessible region
TF binding
Di-nucleosome
Mono-nucleosome
Tri-nucleosome
Tetra-nucleosome
SCA-seq segment count
Length of SCA-seq inaccessible region (bp)
j
All CTCF site
Accessible region
Inaccessible region
CTCF binding
CTCF site with footprint
GpC methylation ratio
-1k
CTCF motif
1k
k
Cardinality distribution
Wilcoxon, p = 0.039
Cardinality
Percentage of concatemer
Type of concatemer
CTCF site with footprint
CTCF site without footprint
Whole genome
l
CTCF site without footprint
CTCF site with footprint
Whole genome
Contact frequency (log10)
Log10 linear distance (bp)

Fig 3. SCA-seq captures higher-order chromatin structure and CTCF footprints. **(a)** Contact frequency (y-axis) as a function of linear genomic distance (x-axis) was plotted across all hg19 chromosomes for SCA-seq (red) and Hi-C (brown). **(b)** Comparison of 1.1 M SCA-seq virtual pairwise contacts (lower triangle) and 1.8 M Hi-C contacts (upper triangle) for chromosomes 1–22 and X (hg19) in HEK293T cells. **(c)** and **(e)** Comparison of SCA and Hi-C 250 kb (c) and 25 kb (e) contact maps for chromosome 7. The corresponding eigenvector (c) and insulation score (e) indicate visual consistency between SCA-seq (red) and Hi-C (brown) signal patterns. Color scale bar: log normalized read counts. **(d)** and **(f)** Scatterplots comparing eigenvector (d) and insulation score (f) between SCA-seq and Hi-C. PCC: Pearson correlation coefficient ($P < 2.6 \times 10^{-16}$). **(g)** Contact map showing an example of CTCF peaks at 10 kb resolution with a loop anchor signal indicated by black circles. Color scale bar: log normalized read counts. The CTCF ChIP-seq tract is plotted in between. **(h)** Aggregate peak analysis showing visual correspondence of enrichment patterns between Hi-C and SCA-seq within 100 kb of loop anchors at 10 kb resolution. Color scale bar: sum of contacts detected across the entire loop sets at CTCF sites in a coordinate system centered around each loop anchor. **(i)** The lower density plot shows the inaccessible region length distribution of SCA-seq. The upper panel shows the schematic diagram of the transcription factor binding and nucleosome footprint patterns. **(j)** The heatmap of GpC methylation level shows chromatin accessibility at the CTCF motif at a single molecule resolution. A subset of CTCF sites with CTCF binding footprint (50 bp accessible–50 bp inaccessible–50 bp accessible) is shown in a magnified view. **(k)** Cardinality distribution of three sets of concatemers. Concatemers with a CTCF footprint (blue) and without the footprint (yellow) are determined by the CTCF binding footprint. The genome-wide concatemers are shown in red. Distributions of concatemers with and without CTCF footprint were compared by the Wilcoxon test. **(l)** Multiplex concatemers were converted to pairwise contacts. The linear distances between pairwise contacts are plotted as density distributions.

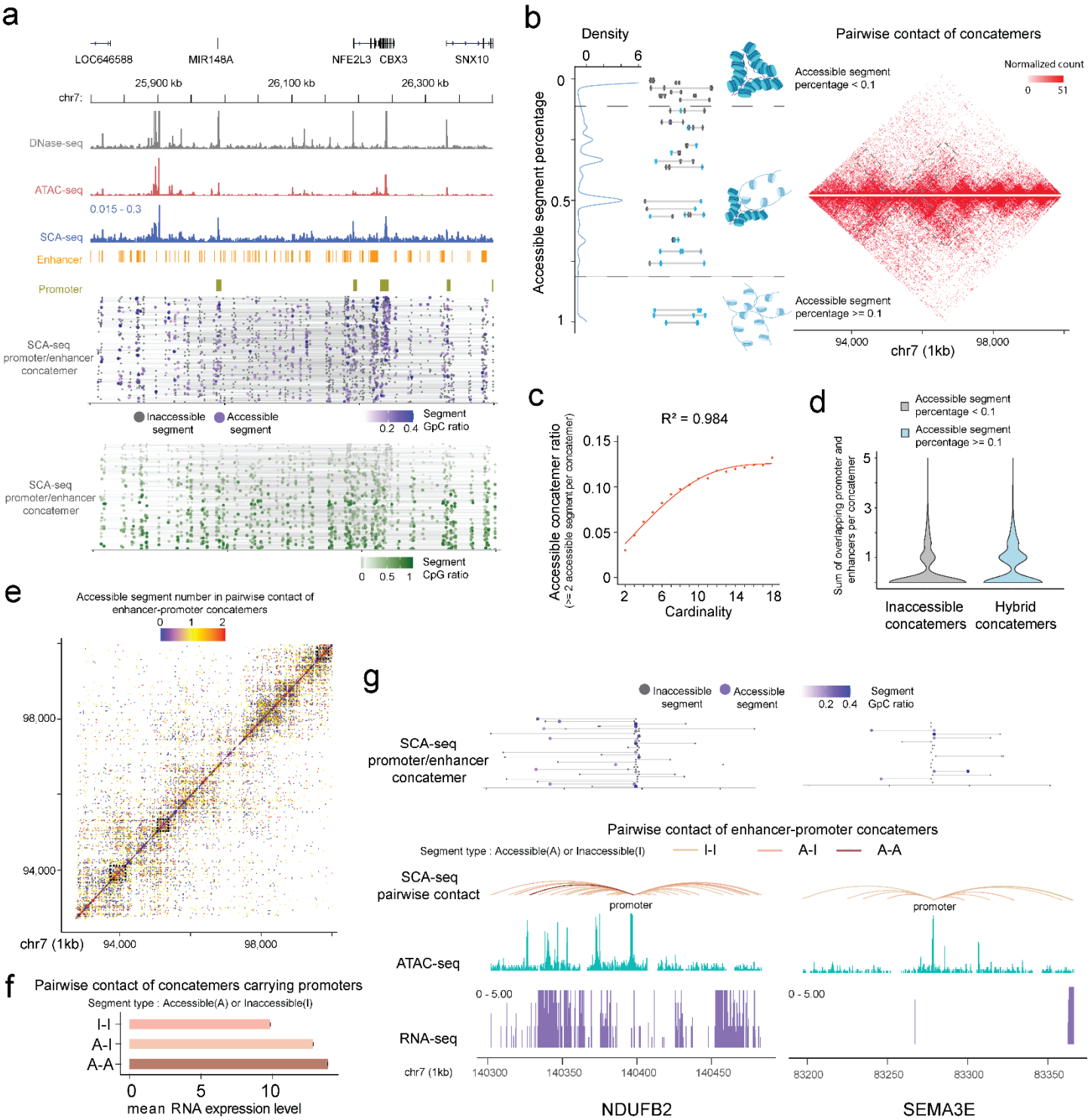


Fig 4. SCA-seq demonstrates spatial chromatin accessibility. **(a)** Demonstration of promoter-enhancer concatemers in the region from 25,800 kb to 26,400 kb on chromosome 7 on a single molecule level. Gray, red, and blue tracks indicate chromatin accessibility signals from DNase-seq, ATAC-seq, and SCA-seq, respectively. The spatially inaccessible segments that were proximately ligated together to form one concatemer are shown as a single line in the bottom panel. Purple and

green dots indicate chromatin accessibility (GpC) and CpG methylation ratio in one segment of a concatemer, respectively. Each row represents a SCA-seq concatemer. **(b)** Distribution of concatemers with specific accessible segment ratios. Each segment was determined accessible or inaccessible. Then the ratio of accessible segments was calculated as follows: $N_{accessible}/(N_{accessible+inaccessible})$. The colored line-dot plot shows GpC methylation of segments (dots) in each concatemer (each line). In total, 150,000 pairwise contacts were subsampled for the different types of concatemers in the genome region chr7:92,700,000–100,000,000 with a 1 kb resolution. Inaccessible concatemers tend to accumulate inside the TAD and hybrid concatemers have more TAD boundary interactions and distal interactions (black box). **(c)** Dot plot with regression curve shows that accessible concatemers (accessible contacts >2 per concatemer) tend to correlate positively with concatemer cardinality (segment number in each concatemer). **(d)** Violin plot of the enhancer/promoter counts in each concatemer. Hybrid concatemers have significantly more enhancers/promoters than inaccessible concatemers ($P < 2.6 \times 10^{-16}$, Wilcoxon rank sum test). **(e)** Accessibility contact map of enhancer–promoter contacts. Inaccessible–inaccessible (I–I), inaccessible–accessible (I–A), and accessible–accessible (A–A) pairwise contacts are shown in blue, yellow, and red colors, respectively. The genome region 92,700,000–100,000,000 is shown at a 1 kb resolution as an example. **(f)** Bar plot showing the relationship of gene expression with different types of accessibility contacts. The y-axis indicates the A–I contact status. The x-axis is the mean RNA expression level under the corresponding promoters (2 kb upstream of the gene). The accessible enhancer/promoter contacts significantly enhanced gene expression ($P < 2.6 \times 10^{-16}$, Student's *t-test*). **(g)** An example of two genes with similar accessibility on the promoter (2.1% accessibility, promoter-centered plot) but different number of A–A contacts. The curve line track is the chromatin interaction detected by SCA-seq. The second and third tracks are ATAC-seq and RNA-seq by counts on the genome. The top line-dot plot shows SCA-seq results at a single concatemer resolution and high-order interactions. The *NDUFB2* gene, with a greater number of accessible contacts, had a notably stronger expression level (528 RPKM) than *SEMA3E* (0 RPKM), which had few accessible contacts.

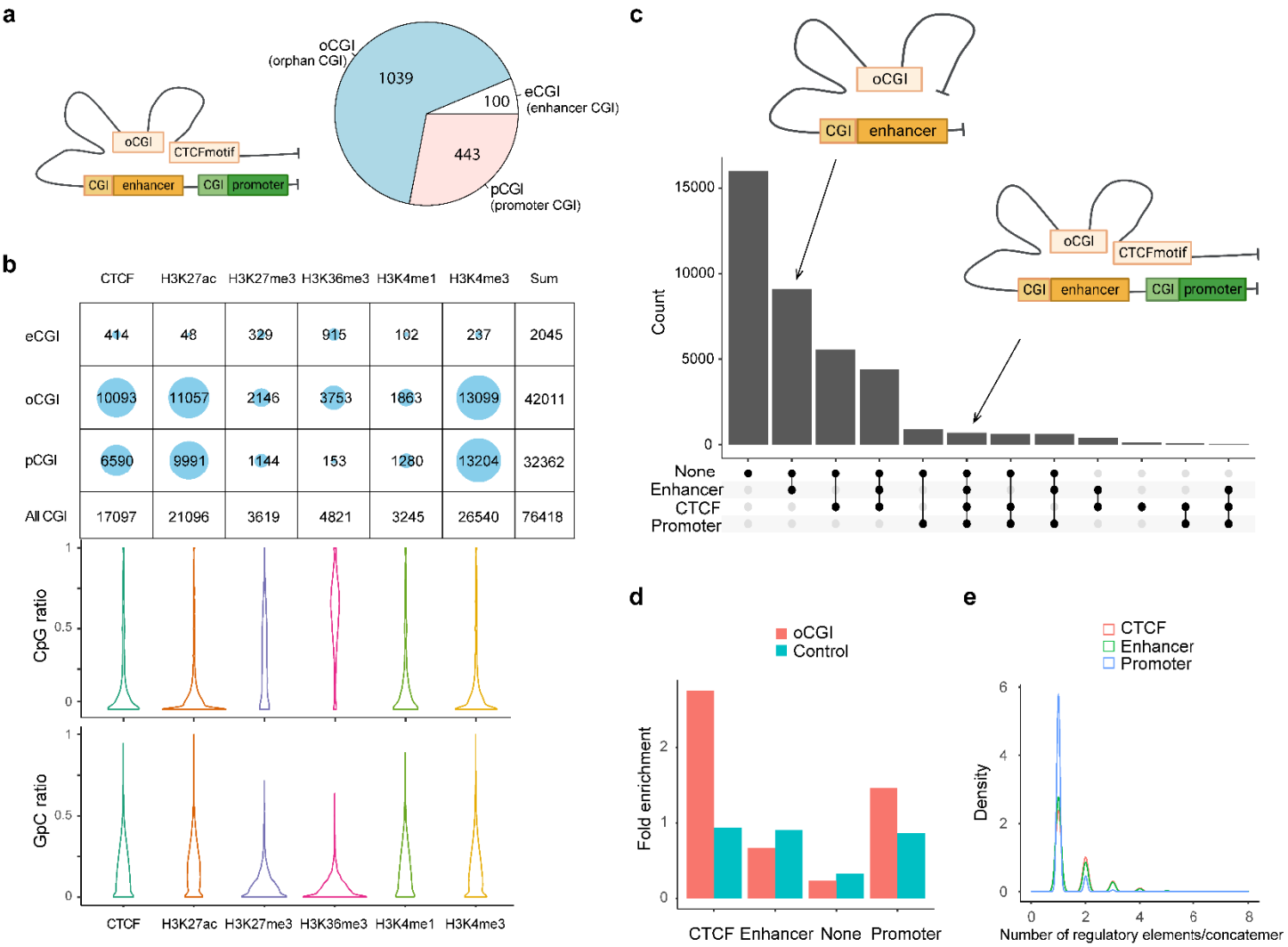
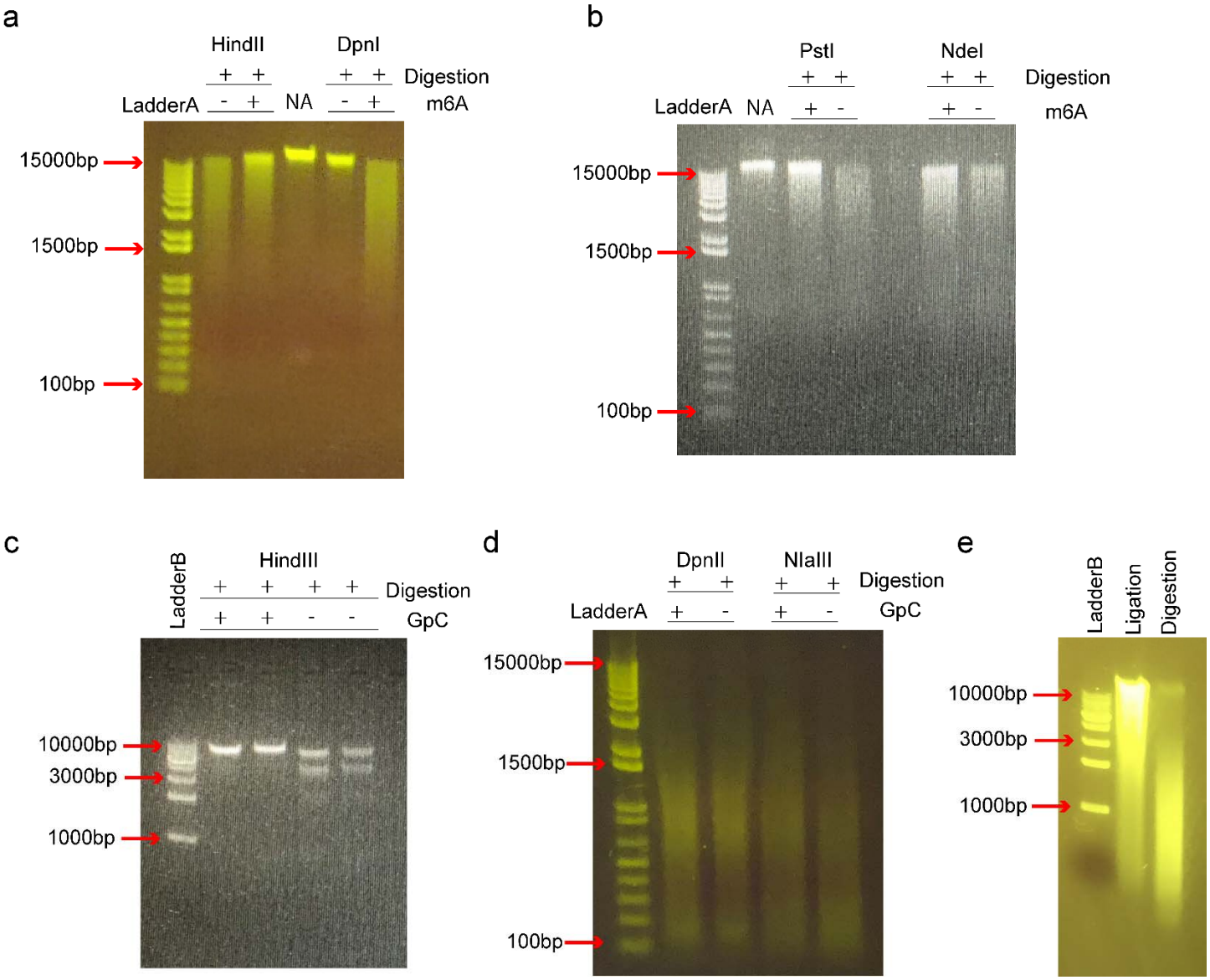


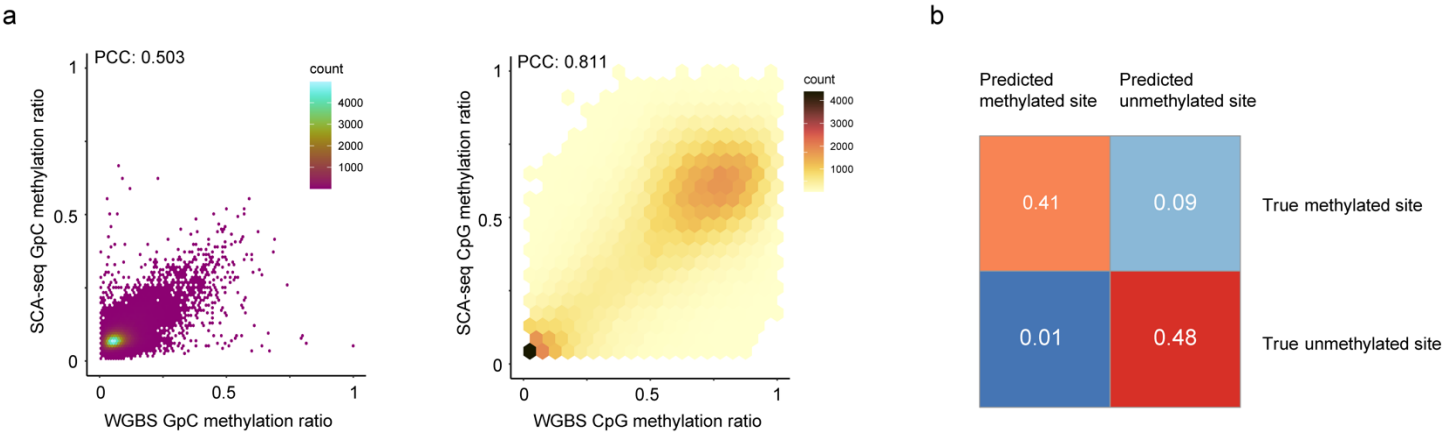
Fig 5. CpG methylation and different types of spatial contacts of CpG islands. **(a)** Schematic illustration of orphan CGI (oCGI) spatial contacts with CTCF, promoter, and enhancer. oCGIs are CpG islands that are far away (at least 1 kb) from enhancers or promoters. The enhancer and promoter annotations were downloaded from ENCODE. The eCGIs and pCGIs are defined as CpG islands near or overlapping with the enhancers and promoters. oCGIs comprised 65% CGIs that did not directly contact enhancers or promoters. The pie chart shows the abundance of different types of CGIs. **(b)** Number of overlaps between three types of CGIs containing concatemers and histone marks/CTCF site. About 83% of the reads contained oCGI, mainly located with active histone markers, such as H3K27ac and H3K4me3, as well as the CTCF binding motif. H3K27me3 and H3K36me3, which are inactive histone makers, were on the opposite DNA methylation state. **(c)** The upset plots indicate types of interaction between oCGI and other elements in one concatemer. On

average, one concatemer has five DNA segments. These segments with oCGIs included CTCF binding motifs, enhancers, promoters, and other proteins. The y-axis indicates the number of concatemers. **(d)** Stacked bar plot shows the abundance of each type of elements interacting with concatemers harboring oCGI or random-sampled concatemers (control). The number of interacting elements was normalized by the total number for itself. oCGIs preferentially interacted with CTCF binding motifs and promoters. **(e)** Density plot demonstrating the distribution of the number of elements on concatemers. Thirty-nine percent of concatemers have more than one enhancer and CTCF binding motif with oCGIs. Promoters were in contact only with oCGIs.

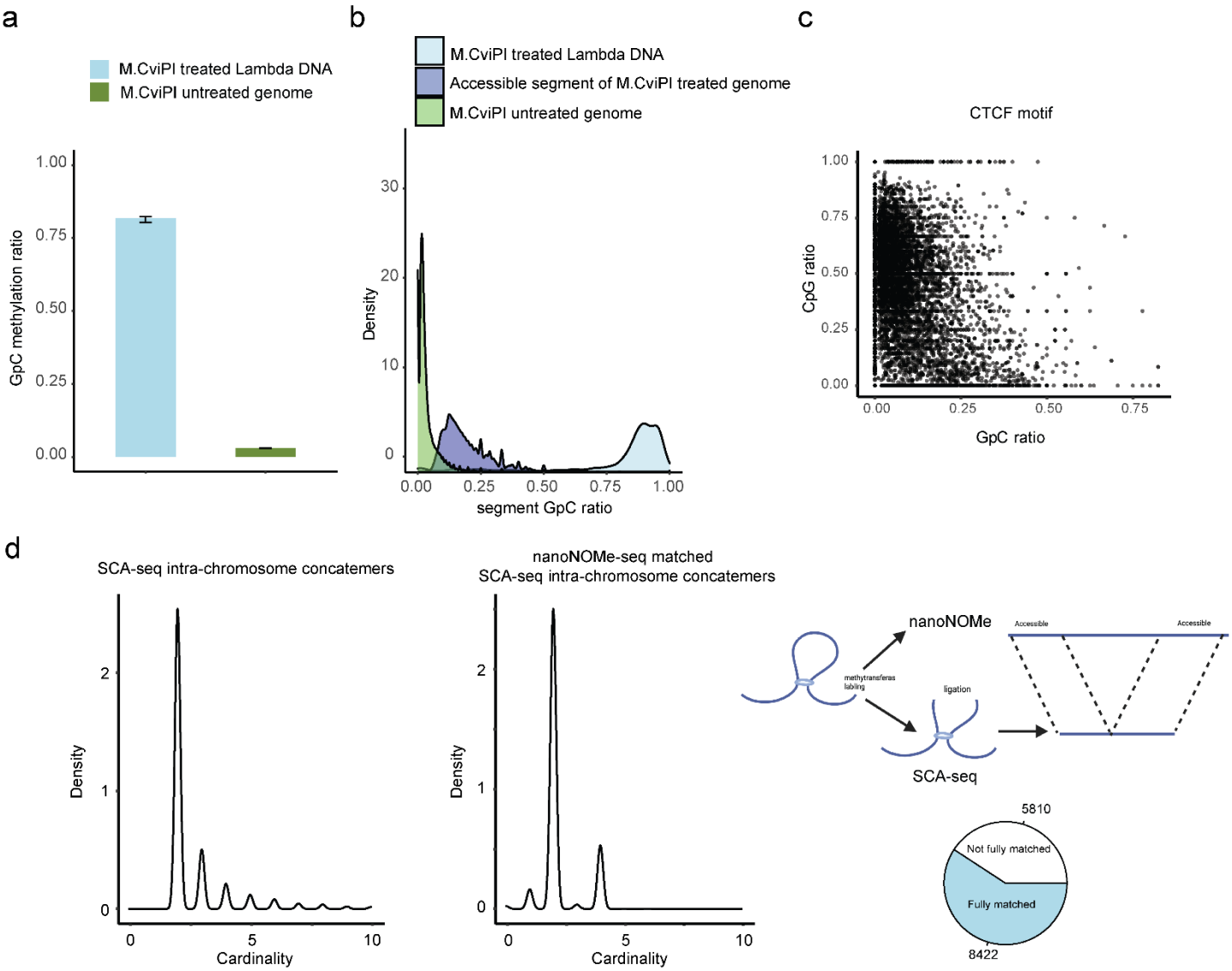


Sfig 1. Selection of the compatible restriction enzymes. In SCA-seq, we used methyltransferases (EcoGII or M.CviPI) to label the accessible chromatin and the restriction enzyme to digest the genome and prepare for the ligation step. Methyltransferase EcoGII transferred the methyl group to the 6′-carbon of adenosine (m6A modification). Due to the high density of adenosines in the genome (~25%), the abundant artificial m6A modification may impair restriction enzyme activity. We used EcoGII to treat the HEK293T genomic DNA ($m6A^+$), whereas control DNA was left untreated ($m6a^-$). We tested HindIII, DpnI **(a)**, PstI and NdeI **(b)** on the methylated and unmethylated genomic DNAs, and all these restriction enzymes were inhibited by the abundant m6A modifications. DpnI is an m6A-dependent restriction enzyme that digests only genomic loci with the m6A modification. NA indicates the genomic DNA control. In a similar way, we tested the activities of HindIII **(c)** for lambda DNA;

DpnII, and NIaIII **(d)** for HEK293T genomic DNA on the GpC methylated DNA preparations that were treated by methyltransferase M.CviPI. DpnII and NIaIII were not significantly inhibited by GpC methylation. **(e)** The HEK293T genomic DNA was digested by DpnII (Digestion) and then ligated by T4 DNA ligase (Ligation).

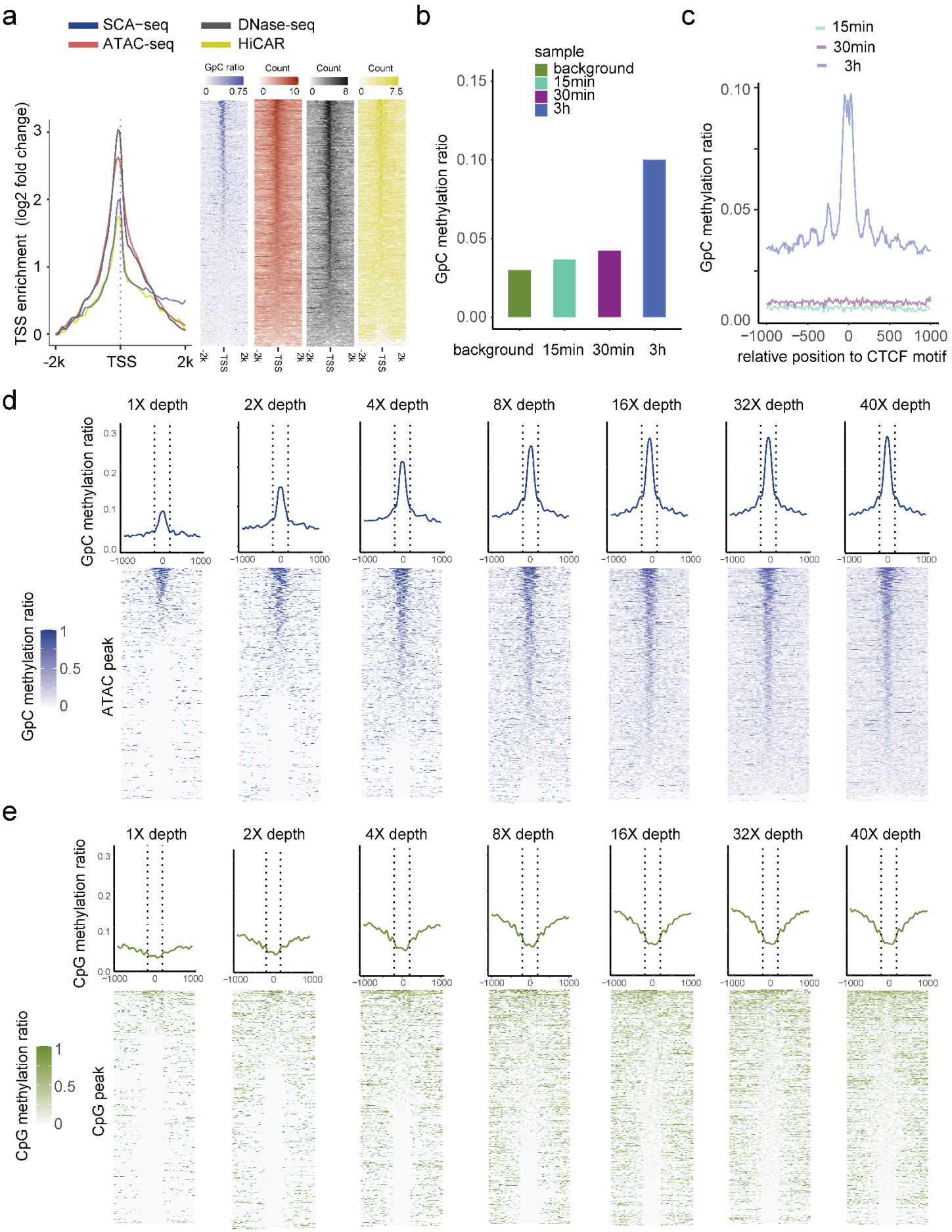


Sfig 2. Feasibility of methylation calling. **(a)** The similarity between the whole genome bisulfite sequencing (WGBS) GpC methylation calls and SCA-seq GpC methylation calls is illustrated. The methylation ratios in 200 bp bins of SCA-seq were scatter plotted against WGBS methylation ratios (Pearson's correlation, $P < 2.6 \times 10^{-16}$). A similar plot of the CpG methylation ratio is on the right (Pearson's correlation, $P < 2.6 \times 10^{-16}$). **(b)** Confusion matrix of GpC methylation calling. Subsets of 100,000 sites in unmethylated pore-C HEK293T samples and methylated lambda DNA samples were methylation called by the default threshold of the Nanopolish cpggpc model. The results are classified into four categories: false positive and true negative from the unmethylated pore-C run and true positive and false negative from the GpC-labeled lambda DNA sample.

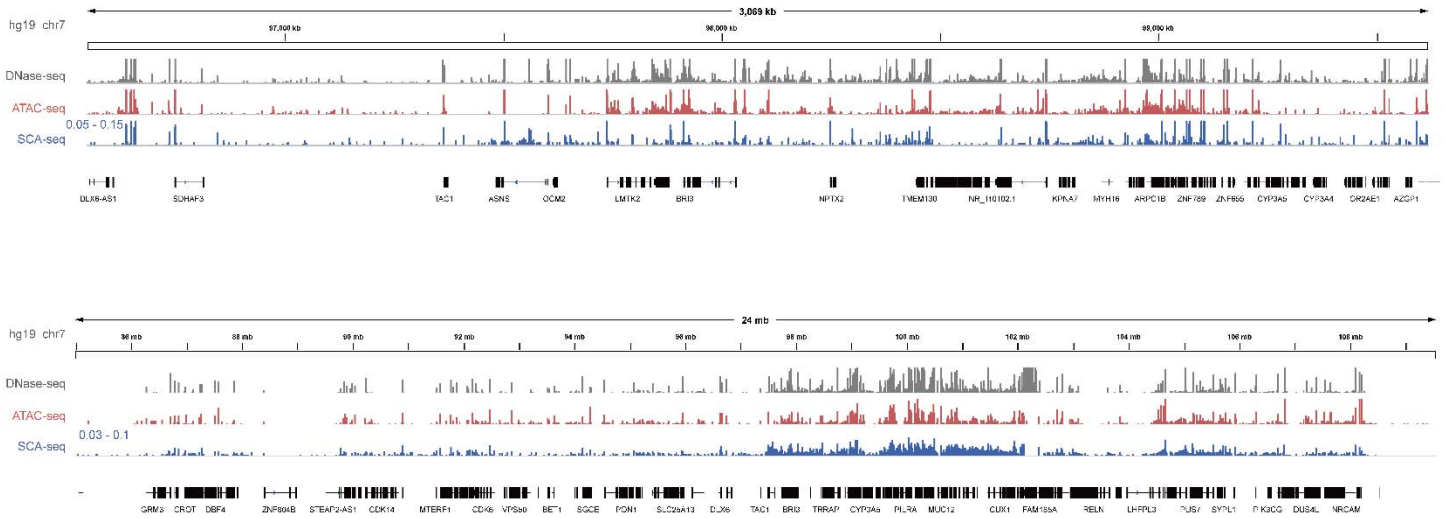


Sfig 3. GpC labeling efficiency *in vitro* and *in vivo*. We used lambda DNA as the spiked-in DNA control to monitor M.CviPI efficiency. Genomic DNA without labeling was used as background control to calculate the background level of GpC. The labeled genomic DNA was processed using SCA-seq. We plotted the density plot of GpC ratios (methylated GpC/total GpC) in each segment for background genomic DNA, treated genomic DNA, and spike-in control DNA. The background GpC ratio in each segment was around 0.018 **(a, b)**. Efficiency of spiked-in lambda DNA labeling by M.CviPI was around 0.88 *in vitro*. We then selected CTCF motifs as control to calculate *in vivo* labeling efficiency **(c)**. We plotted the CpG methylation ratio (methylated CpG/total CpG) against the GpC methylation ratio in CTCF motifs. Previous publications showed that pre-existing methylation could antagonize CTCF binding to CTCF motifs [1-3]. Therefore, we assumed that CTCF motifs with low

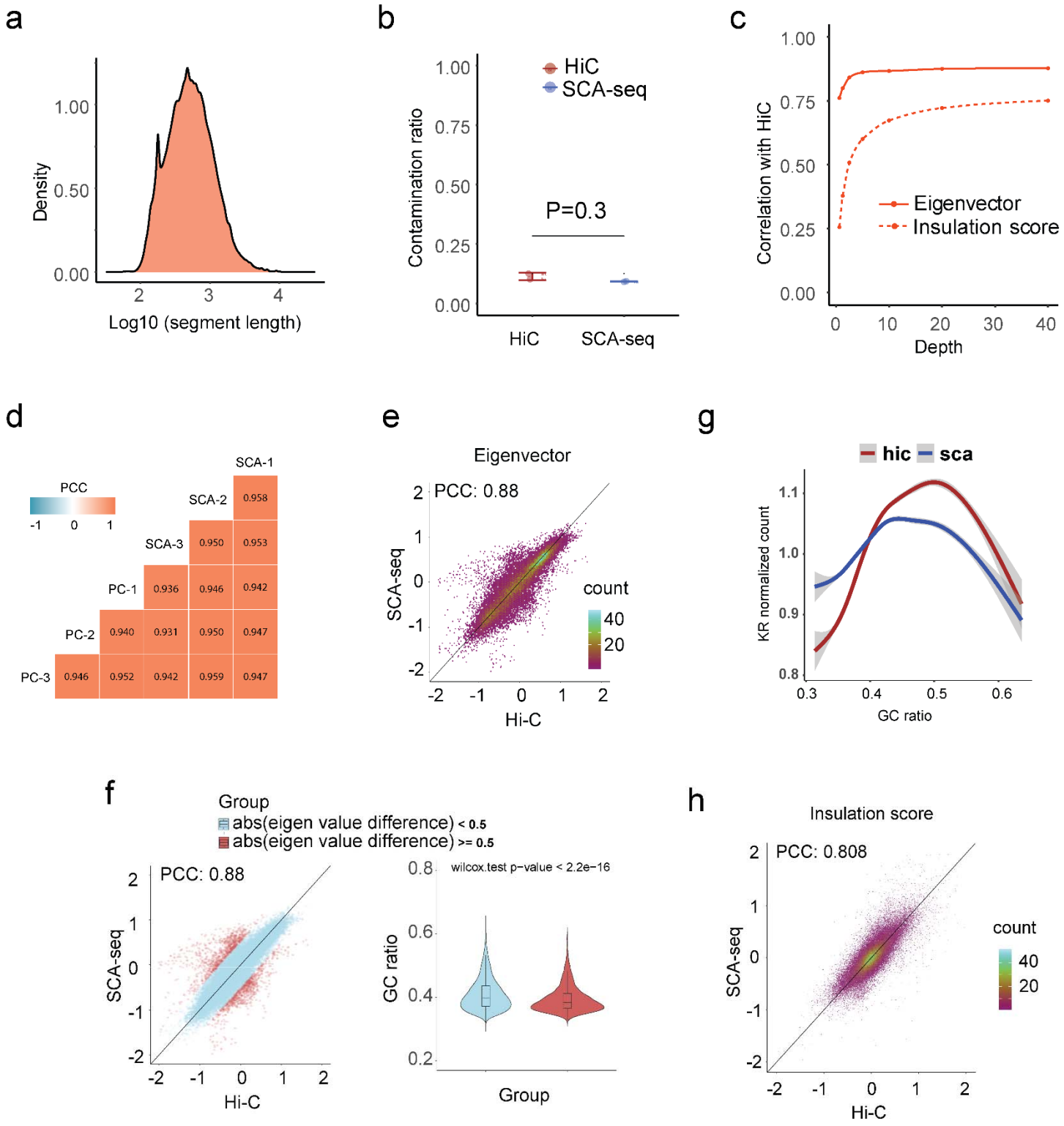
methylation (CpG methylation < 0.25) marked accessible chromatin regions with capacity for CTCF binding. The CTCF motif labeling efficiency was calculated as $N_{CpGmethylation<0.25\&GpCmethylation>0.07}/N_{GpCmethylation<0.25}$=0.79. However, the highly methylated CTCF motifs might also bind CTCF to maintain chromatin accessibility. Therefore, the *in vivo* labeling efficiency could be estimated only roughly. We then used nanoNOME-seq to cross-validate SCA-seq labeling efficiency **(d)**. Both nanoNOME-seq and SCA-seq were single molecule assays. nanoNOME-seq could report chromatin accessibility in 2−20 kb single DNA molecules. Sixty-eight percent of SCA-seq concatemers (d, left panel) corresponded to similar DNAs identified by nanoNOME-seq (d, right panel). The most matched concatemers contained only two segments, representing the short distance interaction. The concatemers with multiple segments, representing the long-distance interaction, may be out of the nanoNOME-seq detection range.


a
SCA-seq
ATAC-seq
DNase-seq
HiCAR
GpC ratio
Count
Count
Count
TSS enrichment (log2 fold change)
-2k
TSS
2k
b
sample
background
15min
30min
3h
GpC methylation ratio
background
15min
30min
3h
c
15min
30min
3h
GpC methylation ratio
relative position to CTCF motif
d
1X depth
2X depth
4X depth
8X depth
16X depth
32X depth
40X depth
GpC methylation ratio
ATAC peak
e
1X depth
2X depth
4X depth
8X depth
16X depth
32X depth
40X depth
CpG methylation ratio
CpG peak

Sfig 4. Dependence of labeling efficiency and signal strength from enzyme concentration and sequencing depth. **(a)** TSS-centered plot illustrating the enrichment score (Y-axis) for different methods around TSSs (0 on the X-axis), which shows a significant nucleosome depletion pattern. The enrichment score was calculated as log2(signal in each bin/lowest signal in scale). **(b)** Test of labeling efficiency at three different labeling times (15 min, twofold higher enzyme concentration; 30 min, twofold higher enzyme concentration, 3 h, threefold higher enzyme concentration). GpC labeling efficiency was the highest after the 3 h treatment, and the labeled GpC signal accurately showed the CTCF motif nucleosome pattern **(c)**. The results of the 3 h treatment were also similar to those obtained for nanoNOME-seq tissue sample preparation [4]. **(d, e)** ATAC-seq peaks were identified first, and methylation ratios obtained using SCA-seq are plotted around the identified peaks (center to 0) with different sequencing depths. The line panels show a single molecular resolution.

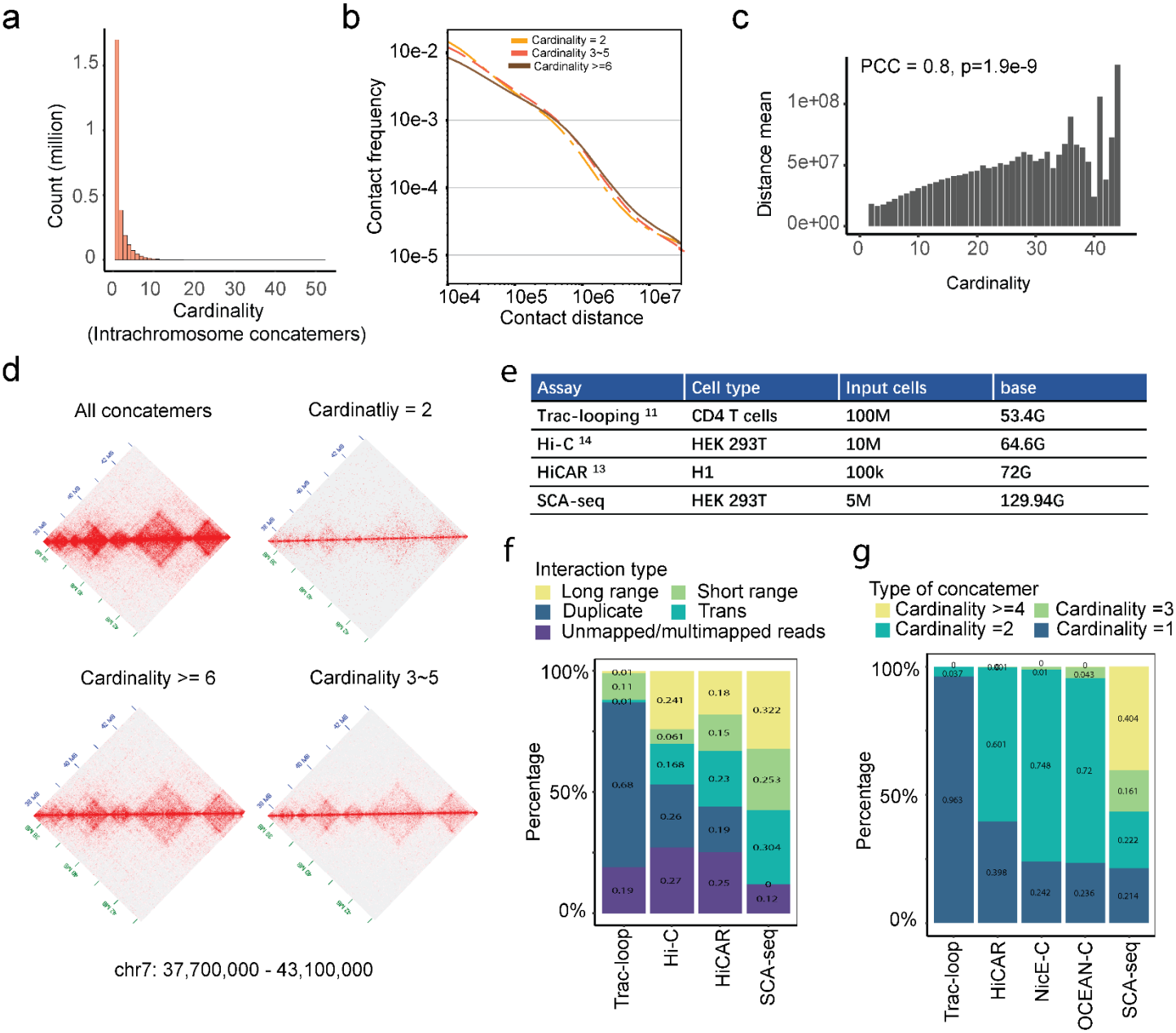


Sfig 5. Chromatin accessibility at various resolutions. The panels show DNase-seq, ATAC-seq, and SCA-seq raw data in the genome browser. DNase-seq and ATAC-seq results show read counts in each genomic locus, whereas SCA-seq data represent methylation site counts in each genomic locus.



Sfig 6. Quality parameters of SCA-seq for resolving the genome structure. **(a)** The medium length of segments in concatemers was around 700 bp. **(b)** We further calculated the false positive contacts caused by the ligations (contamination). We used the contacts between the mitochondrial genome and nucleic genome/total contacts to calculate the contamination ratios. SCA-seq and Hi-C contamination ratios were 9% and 12%, respectively. **(c)** We then subsampled the reads to test the

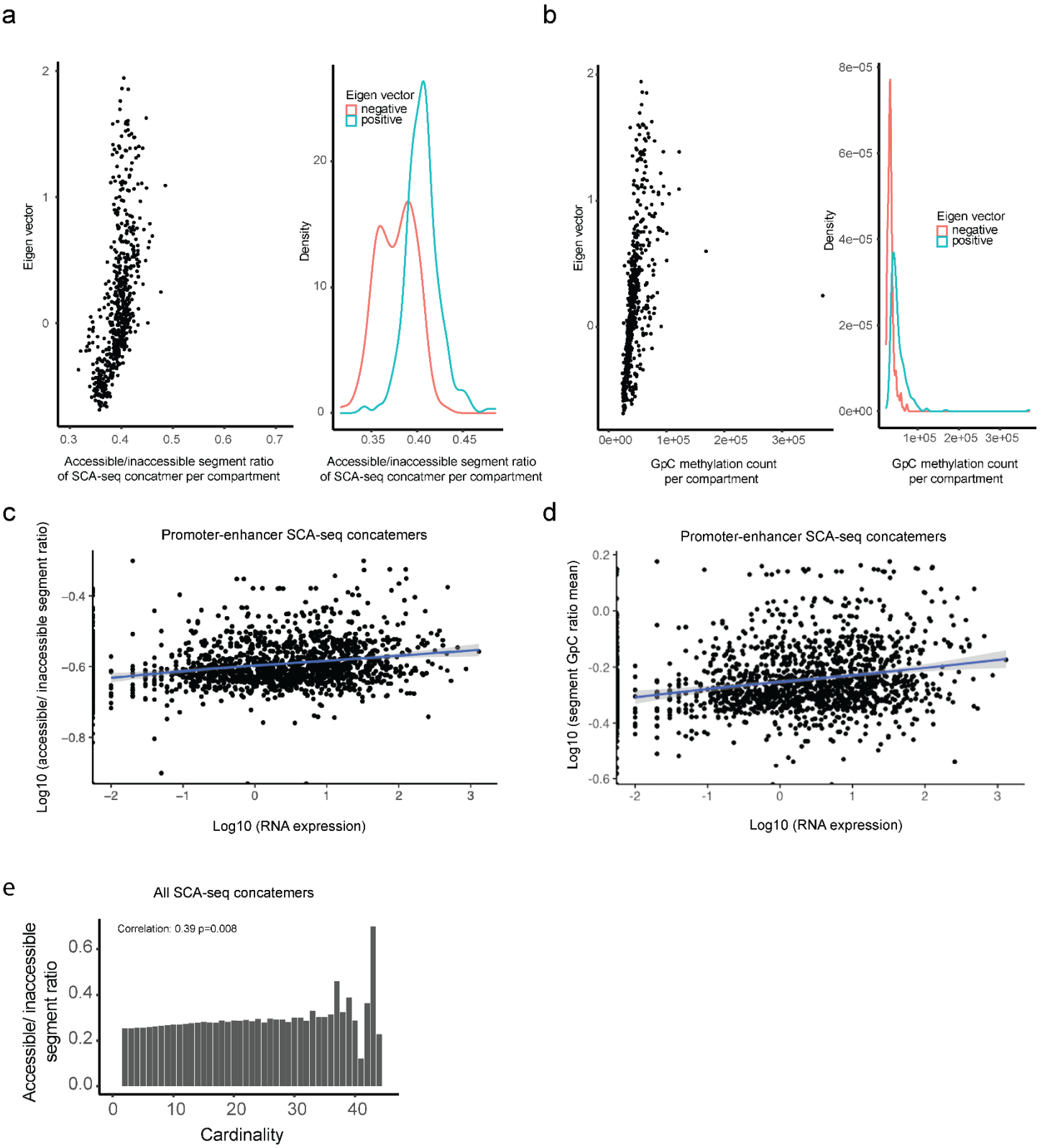
sequencing threshold to calculate the eigenvector (AB compartment) and insulation score (TAD). We needed over 5 million and 30 million reads to accurately produce results similar to the Hi-C output for in the AB compartment and TAD, respectively. **(d)** We then tested the correlation (0.93–0.94) between SCA-seq and pore-C replicates [5]. SCA-seq demonstrated good consistency between samples. **(e)** Plot of global eigenvalue correlation (without quality control) between SCA-seq and Hi-C data. **(f)** The compartment correlations with bias were far from the diagonal line and are colored red. The bias regions had a significantly lower GC ratio than the unbiased region in the violin plot. **(g)** The GC ratio distribution in SCA-seq and Hi-C. **(h)** Plot of global insulation score correlation (without quality control) between SCA-seq and Hi-C data.



Assay	Cell type	Input cells	base
Trac-looping [11]	CD4 T cells	100M	53.4G
Hi-C [14]	HEK 293T	10M	64.6G
HiCAR [13]	H1	100k	72G
SCA-seq	HEK 293T	5M	129.94G

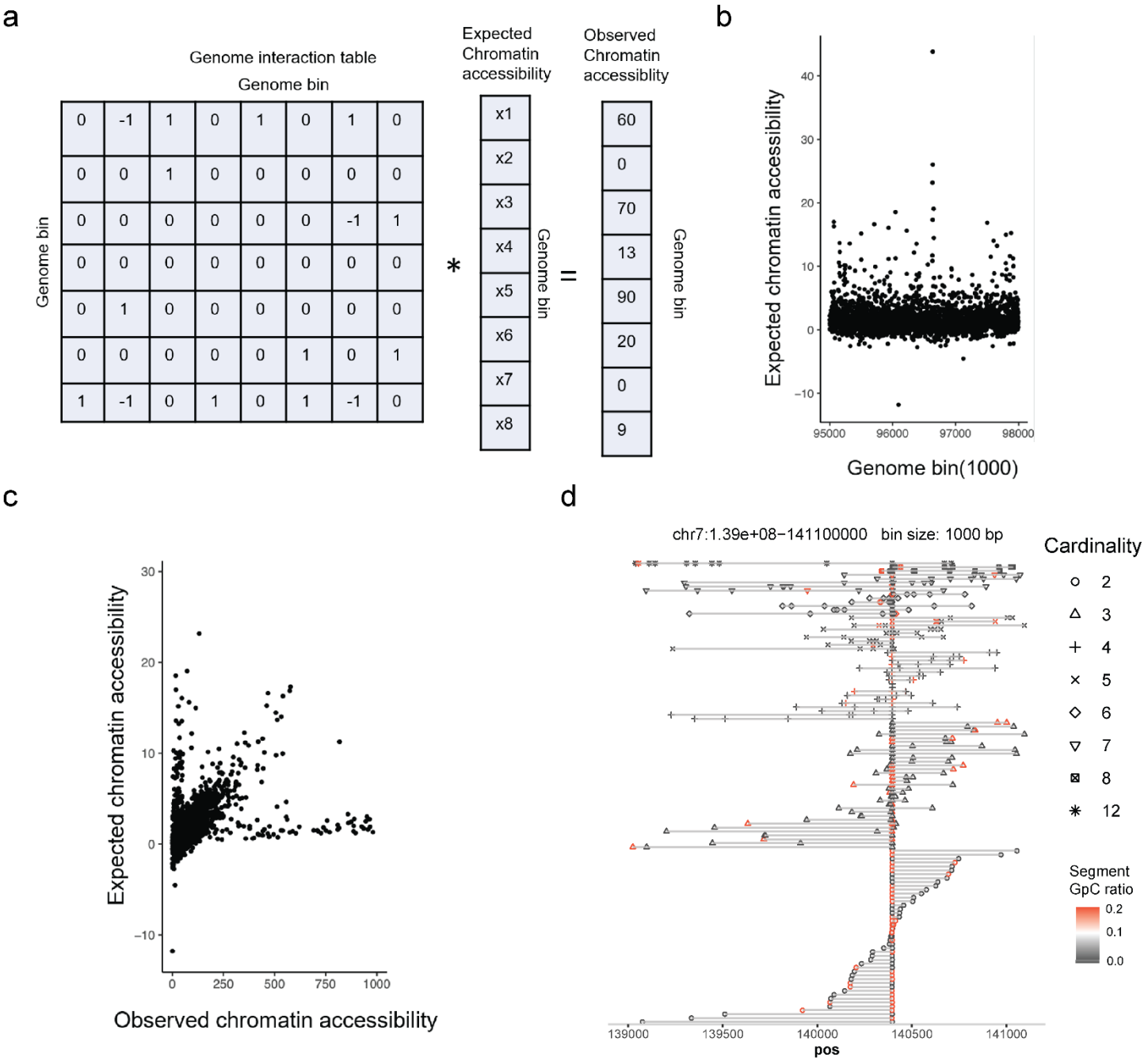
Sfig 7. High-order contacts resolve the long-distance genome interaction. **(a)** Intra-chromosome contacts on chromosome 7. Sixty-five percent of concatemers contain one segment without any genome contacts, 14.7% of concatemers contain two segments, 14.5% contain 3–5 segments. 5.4% of concatemers have ≥ six segments. **(b)** Plot of the contact frequency against the contact distance for concatemers with different cardinality values. The high-cardinality concatemers had more distant contacts than the low-cardinality concatemers. **(c)** The maximum contact distance in each concatemer positively correlates with concatemer cardinality. **(d)** Concatemers with different cardinality values were converted to pairwise contact maps and normalized by the JuicerTools KR

method. The high-cardinality concatemers tend to maintain the TAD structures and indicate long-distance interaction. **(e)** The data, cell type, and cell number of each method. **(f)** Interactions at distances of ≥ 20 kb and < 20 kb were classified as long-range and short-range interactions. “Duplicate” indicates duplicated reads. “Trans” indicates interaction across chromosomes. **(g)** The cardinality represented the segment number in one read, referring the conformation complexity.



Sfig 8. Correlation between compartmental eigenvalue, RNA expression, and spatial accessibility. **(a)** Plot of compartmental eigenvalues against the average concatemer accessible/inaccessible ratio in the compartments. The average concatemer accessible/inaccessible ratio was correlated with the

compartmental eigenvalue (Pearson correlation; $P < 2.6 \times 10^{-16}$). The average concatemer accessible/inaccessible ratio was significantly higher in the A compartments (positive eigenvalue) than in the B compartments (negative eigenvalue) ($P < 2.6 \times 10^{-16}$, Student's *t-test*). **(b)** High correlation of compartmental eigenvalues with labeled GpC methylations. The labeled GpC methylations were highly correlated with the compartmental eigenvalue (Pearson correlation; $P < 2.6 \times 10^{-16}$). The labeled GpC methylations were significantly more abundant in the A compartments (positive eigenvalue) than in the B compartments (negative eigenvalue) ($P < 2.6 \times 10^{-16}$, Student's *t-test*). **(c)** For enhancer/promoter accessibility contacts in Fig 4d, we summarized the mean contactable accessibility on the promoters/enhancers (inaccessible/inaccessible 0, inaccessible/accessible 1, accessible/accessible 2). The mean contactable accessibilities on promoters highly correlated with the expression level of the corresponding gene (Pearson correlation, $P < 2.6 \times 10^{-16}$). **(d)** Similarly, we summarized the mean GpC methylation ratio on the concatemers with promoter/enhancer regions. Mean concatemer GpC methylation ratios on promoters/enhancers significantly correlated with corresponding gene expression levels (Pearson correlation, $P < 2.6 \times 10^{-16}$). **(e)** The accessible/inaccessible ratios of concatemers were positively correlated with the cardinality (Pearson correlation, $P = 0.0008$). As shown in Sfig 6 b,c, we found that high order concatemers indicate long-distance interactions.



Sfig 9. Expected chromatin accessibility. A chromatin region could be made accessible by the contacts with a spatially inaccessible genome locus. We designed an equation to eliminate the effects of the genome interaction. **(a)** We used SCA-seq to compile a table of interactions between genome bins. In this genome interaction table, the accessible–accessible contacts were defined as 2, which means the contacts could initiate the activation. The inaccessible–inaccessible contacts were defined as −2, and the accessible–inaccessible contacts were defined as 0. Then these values were summarized in the genome interaction table as means. The 3D genome interaction matrix times the

expected accessibility was equal to the observed chromatin accessibility. Then we resolved the matrix equation. The expected chromatin accessibility represented the original chromatin accessibility without the genome interaction effects. The high signal in expected chromatin accessibility also indicated that this site, which we recognized as the “power center”, could potentially activate the contacting sites. **(b)** The dot plot of expected chromatin accessibility for a region of the genome. **(c)** Dot plot of the expected chromatin accessibility (y-axis) against the observed chromatin accessibility (x-axis). **(d)** Selection of a representative CTCF motif with highly expected signals. Each line represents a concatemer segment in the representative genome region.

1. Bell, A.C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**, 482-5 (2000).
2. Hark, A.T. et al. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**, 486-9 (2000).
3. Kanduri, C. et al. Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. *Curr Biol* **10**, 853-6 (2000).
4. Lee, I. et al. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nature Methods* **17**, 1191-1199 (2020).
5. Ulahannan, N. et al. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. *bioRxiv*, 833590 (2019).