Spatial chromatin accessibility sequencing resolves next-generation genome architecture

Yeming Xie[1,†],Yaning Li[1,†], Fengying Ruan[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], 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

[†]These authors contributed equally to this work.

Keywords: HiC, chromatin accessibility, open chromatin, close chromatin, interaction, methylation, m6A, GpC, CpG

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

[*]Correspondence:

Chong Tang

Director of Technology, BGI-Shenzhen, China

Phone: +8618025420976

Email: tangchong@bgi.com

Abstract

As the genome has a three-dimensional structure in intracellular space, epigenomic information also has a complex spatial arrangement. However, the majority of epigenetic studies describe locations of methylation marks, chromatin accessibility regions, and histone modifications in the linear dimension. Proper spatial epigenomic information has never been obtained. In this study, we designed spatial chromatin accessibility sequencing (SCA-seq) to reveal the three-dimensional map of chromatin accessibility and simultaneously capture the genome conformation. Using SCA-seq, we disclosed spatial regulation of chromatin accessibility (e.g. enhancer-promoter contacts) and spatial insulating functions of the CCCTC-binding factor. We demonstrate that SCA-seq paves the way to explore epigenomic information in the three-dimensional space, and extends our knowledge in genome architecture.

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. Accessibility of chromatin acts as a potent gene expression regulatory mechanism by preventing access of regulatory factors to condensed chromatin domains. Although this model is attractive, it is simplified in that the genome accessibility is considered only in the linear dimension[1]. However, the genome has 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. Therefore, sophisticated tools are necessary to obtain information about genome accessibility in three dimensions to resolve the relationship between chromatin activation and genome structure.

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, Song and Crawford used DNase-seq to establish the relationship between DNase-hypersensitive regions and open chromatin[2]. MNase-seq relies on a similar concept[3]. Subsequent studies simplified experiments on chromatin accessibility by taking advantage of the ability of mutant transposase to insert sequencing adapters into open chromatin domains[4]. All these methods rely on statistical calculations of chromatin domain accessibility based on the frequency of attacking events of the enzymes on the accessible genome. To understand the heterogeneity of chromatin accessibility in vivo, scientists have been interested in chromatin structure at a single-molecule resolution level in recent years. They developed approaches such as methyltransferase treatment followed by single-molecule long-read sequencing[5], single-molecule adenine methylated oligonucleosome sequencing assay[6], nanopore sequencing of nucleosome occupancy and methylome[7], single-molecule long-read accessible chromatin mapping sequencing[8 9], and fiber-seq[10]. Subsequently, decondensed genomes were methylated using methyltransferases and then 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 linearized two-dimensional map of chromatin accessibility does not fully reflect the reality. Some advanced approaches, such as Trac-looping[11], OCEAN-C[12], and HiCAR[13], could decrease the need for Hi-C data[14] by enrichment of open chromatin regions by combining transposons and proximity ligation. However, the losses at the single-molecule resolution level and the condensed domains restrict the possibility to observe dynamic changes in chromatin structure in three-dimensional space. Therefore, reconstructing chromatin spatial accessibility could promote further understanding of the interactive regulation of transcription and enable more spatially realistic studies of the genome.

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 and chromatin conformation information were sequenced using nanopore technology. We mapped chromatin accessibility to genome spatial contacts, and the 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 structure.

Results

Principle of SCA-seq

Recently, there has been an increasing interest in applying methylation labeling and nanopore sequencing for analysis of chromatin accessibility at the single-molecule level[5-10]. In this study, we have developed SCA-seq to study chromatin spatial density by updating the 2D chromatin accessibility map to the three-dimensional space (Fig.1a). After cell fixation, we used a methyltransferase enzyme (EcoGII or M.CviPI) to artificially mark “open” chromatin regions.

After the two-dimensional chromatin accessibility information was preserved as methylation marks, we conducted digestion and ligation steps using chromatin conformation capturing protocols, relying on proximity ligation to suture together multiple linearly distant DNA fragments that happen to be close to each other in three-dimensional space. The DNA fragments that carried both two-dimensional chromatin accessibility marks and three-dimensional conformation information were sequenced using the nanopore method and analyzed in our house pipeline (Fig. 1b).

SCA-seq captured multiple aspects of chromatin information, including three-dimensional chromatin conformation, chromatin accessibility, and native CpG methylation on the same segment. Unlike the conventional Hi-C, the proximity ligation, 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-order genome conformation. Moreover, the novel spatial chromosome conformations contain information about spatial interactions among euchromatin, heterochromatin, and hybrid chromatin regions (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[15 16]. Our previous research showed that EcoGII effectively labels open chromatin owing to the high density of adenine the genome[9]. However, the high-density m6A modification either blocked or impaired the activity of the restriction enzymes (Sfig 1a, b). 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, the m6A-dependent digestion

generated biased digestion of the highly methylated open chromatin, and the blunt ends were not ligated efficiently. We then tried another approach and used M.CviPI that methylates GpCs(m5C) on the open 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 both GpC methylated and unmethylated DNAs (Sfig 1c–e).

It should be noted that the m5C base-calling algorithm has been gradually improved and is now widely used in nanopore sequencing[17]. 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 accurately identifies accessible chromatin and methylation marks at single-molecule resolution in two-dimensional space

Our work was based on the concepts of nanoNOME-seq, SMAC-seq, and Fiber-seq[5-10], which use either M.CviPI or EcoGII methyltransferases to label chromatin accessible regions with methylation sites. Our previous experiments[9] 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. The single-molecule solution and methylation caller in this study were identical to those previously described[7, 8].

First, we performed initial quality control of the sequencing data for the HEK293 cell line. We generated 129.94 Gb (36.9× coverage) of mapped sequencing data with an N50 read length of 4,446 bp. We adopted the modified Nanopolish approach[7] as a methylation caller with considerable success (AUC CpG = 0.908, GpC = 0.984), as others had successfully used it for

GpC/CpG calling. The parallel whole-genome bisulfite sequencing data correlated very well with the results of methylation calling (Sfig 2), further supporting our choice of the methylation caller. In the analysis, we excluded both CpG and GpC methylation data from the GCG context because of the ambiguity of native methylation or chromatin state labels (5.6% of GpCs and 24.2% of CpGs). The native or false-positive GpC regions were very rare and only accounted for 1.8% of the control genome (Sfig 7a, b).

We next assessed the potential of SCA-seq to reveal simultaneously endogenous methylation and chromatin accessibility. SCA-seq data correlated with those obtained using the assay for transposase-accessible chromatin (ATAC)-seq and DNase-seq at various resolution levels (Fig 2e; Sfig 3). Of the 2,995 SCA-seq data peaks on chromosome 7, 72% overlapped with those observed after ATAC-seq and DNase-seq (Fig 2a). The signal correlation in common peaks was approximately 0.5. Moreover, we used computationally predicted binding sites of the CCCTC-binding factor (CTCF), which are a well-documented open chromatin indicator,[18] and our data were supported by those of CTCF ChIP-seq. CpG methylation level decreased, and GpC accessibility increased around the CTCF-binding sites (Fig 2b). SCA-seq also showed peak patterns around the ATAC-seq identified peaks (Fig 2c). At active human transcription start sites (TSSs), “open” chromatin regions hypersensitive to transposon attack were observed. Furthermore, SCA-seq showed similar nucleosome depletion patterns around TSSs (Fig 2d). Inactive TSSs were less accessible than active TSSs (Fig 2d). The relationship between the dose and M.CviPI treatment effect demonstrated superior efficiency of the 3 h treatment, comparing with 15min, 30min treatment (Sfig 4). Overall, SCA-seq reliably estimated chromatin accessibility at the genome level.

SCA-seq reveals high-order chromatin organization

We processed non-singleton chimeric reads into genomic segments and assembled *in silico* paired-end tags (PETs). The segment median length was approximately 700 bp (Sfig 5a). Among the informative PETs, 0.1% of the PETs 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. PETs reveal the high-order chromatin organization, and 60.2% of them reflected intrachromosomal interaction. Unlike Hi-C, SCA-seq, derived from pore-C, revealed the multiplex nature of chromatin interactions: 14.7% of the reads contained two segments; approximately 14.5% of the reads contained 3–5 segments; and 5.4% of the reads had more than five segments (Sfig 6a). As expected, most of the contacts from the reads with low heterozygosity appeared to closely interact. The contacts from the reads with high heterozygosity appeared to have a distal interaction (Sfig 6c, d).

False positivity rates of SCA-seq and HiC inferred from hybrid PETs that consisted of mitochondrial DNA and genomic DNA were similar (Sfig 5b). The compartment score correlation between SCA-seq replicates and pore-C replicates was approximately 0.94 (Sfig 5d). Thirty million reads were enough to resolve the A/B compartment and topologically associating domain (TAD) structures (Sfig 5c).

In further analysis, SCA-seq revealed genome organization similar to the one detected using *in situ* Hi-C. Side-by-side visualization of interaction heatmaps, loops, TAD boundaries, and A/B compartments obtained using SCA-seq and Hi-C showed equivalent genome organizations (Fig 3). Sixty-six percent of the concatemers were case-specific (all the fragments in one

concatemer belonged to A/B compartments) and 34% were non-specific. These results suggested that SCA-seq successfully resolved the multiplex nature of chromatin interactions.

SCA-seq reconstructs chromatin accessibility in three-dimensional space

Given the high cellular heterogeneity in the genome space, our spatial chromatin 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[7]. Besides base calling, we also determined the enzyme labeling efficiency, which was 79–88%, based on the CTCF motifs and spike-in control measurements (Sfig 7a–c). Then, we filtered the fragments using the binomial test to minimize the false positive open/close status (see Methods). Overall genome concatemer calculations showed that 29% of the genome concatemers maintained heterochromatin/closed status on all enclosed fragments (heterochromatin concatemers, open/close chromatin ratio < 0.1). Furthermore, 62.2% of genome concatemers had partial open chromatin fragments (hybrid concatemers), and only 8.8% maintained all chromatin fragments as open (euchromatin concatemers, open/close chromatin > 0.9) (Fig 4a). nanoNOME-seq that labels chromatin accessibility in single molecules also confirmed the existence of these concatemers (Sfig 7d). Chromatin accessibility was related to spatial contacts. For example, the hybrid concatemers tended to gather around the TAD boundary and contain more distant connections (Fig 4b). The heterochromatin concatemers had fewer fragments than the hybrid concatemers, implying their interaction preference (Fig 4c, Sfig8e, Sfig6bc). The A/B compartments are usually related to chromatin accessibility and regions of gene expression [1]. In our study, we found that the B compartment (negative eigenvector) had significantly fewer euchromatin concatemers

(open/close chromatin ratio 0.37 vs. 0.4, $P < 2.2 \times 10^{-16}$) than the A compartment (positive eigenvector) (Fig 4d; Sfig 8a, b). We further investigated the enhancer and promoter contacts on chromosome 7, 30.3% of which had open–open status; 18.5%, open–close status; and 51.2%, close–close status (Fig 4e). Most promoter/enhancer contacts spontaneously initiate the open–open status upon contacting. The frequency of contacts with open chromatin highly correlates with gene expression levels (Fig 4f, g; Sfig 8c, d), supporting the transcription model in which promoters contact enhancers[19]. The heterogenous accessibility pattern within the proximal spatial distance suggests very complex epigenetic regulation of the mammalian genome. The spatial contacts and contact accessibility might coordinately regulate the gene expression.

Spatial chromatin accessibility is dramatically changed in CTCF$^{-/-}$ HEK293T cells

In mammals, the highly conserved zinc finger protein CTCF is thought to serve as an insulator protein that prevents communication between enhancers and promoters and thereby, regulates chromosome folding[20 21 22]. However, it has been shown recently that CTCF loss had minimal effects on global genome architecture in Drosophila[23]. We examined the impact of permanent CTCF loss in HEK293T cells and observed only a minor effect on the global genome architecture and TAD boundary (Sfig 9). However, we and others found that CTCF loss slightly activated genome accessibility[24] (Sfig 10). This was consistent with the lack of significant changes in the two-dimensional GpC methylation ratio (Sfig 10). We further studied whether chromatin accessibility was altered by the spatial contacts. Overall, the number of hybrid/open concatemers increased substantially, and most of them had more than one open chromatin contact (Fig 5a). In the contact maps with accessibility plots, the chromatin spatial

accessibility had a chessboard pattern in wild-type cells. The open/half-open chromatin contacts were arranged in lines on the chessboard, suggesting that the loci contacted specific open chromatin regions, such as CTCF motifs and ATAC peaks (Fig 5b; Sfig 11). In the $CTCF^{-/-}$ cells, chromatin was activated mainly by contacts, and the spatially activated chromatin changed the chessboard/line patterns (Fig 5b; Sfig 11). In the static analysis of the contact map, we found that these activations were very significant and distributed with the CTCF ChIP signal (Fig 5c). The activation folds were much larger outside than they were inside the TADs ($P < 2.6 \times 10^{-16}$, *t*-test) (Fig 5d). The sparser data outside the TADs may bias this finding. However, this observation might provide a clue to the CTCF insulator function conundrum in that CTCF possibly prevents the impropriate chromatin activation or inactivation by the outside-TAD contacts even if CTCF did not physically block these contacts. We found that the loss of CTCF activated the outside-TAD contacts and induced abnormal activation of the inside-TAD contacts (Fig 5b; Sfig 11). As mentioned earlier, CTCF-binding sites maintained open chromatin status, so any accessibility contacts with CTCF motifs were either 1 (open–close) or 2 (open–open). According to our hypothesis, in $CTCF^{-/-}$ cells, the loss of insulation generates more open–open chromatin contacts. In $CTCF^{-/-}$ cells, 29.4% of genomic regions in contact with CTCF motifs had increased chromatin accessibility (open–open contacts)(Fig 5e). In addition, in $CTCF^{-/-}$ cells, the majority of the increased open–open chromatin contacts significantly correlated with the CTCF ChIP signals. This finding further confirmed the close relationship between CTCF loss and spatial activation. Some CTCF loops restrict the ability of active and inactive regions to segregate into compartmental domains. In contrast, other CTCF loops increase the frequency of interactions between two adjacent active and passive domains[22] (Fig 5f, carton). We aggregated CTCF loop signals together and found

the chessboard pattern of spatial chromatin accessibility, suggesting contact-based activation by the CTCF loops. The loss of CTCF greatly enhanced the contact-based activation by the CTCF/cohesion loops ($P < 2 \times 10^{-16}$, *t*-test). Also notable is that chromatin accessibility was significantly enhanced in the center area, at the CTCF loop location ($P < 2 \times 10^{-16}$, *t*-test). Some CTCF loops increased the interaction between adjacent active and inactive domains and insulated each chromatin status. The loss of CTCF might prompt invasion of the idle domain by the adjacent active domain. To quantify CTCF effects, we developed an algorithm to calculate the activation power of the genome loci (Sfig 12). The locus with high activation power might enhance spatially proximal genome accessibility. We found that CTCF loss increased the activation power in most regions without significantly changing the global accessibility. Overall, with SCA-seq, we found that CTCF loss led to contact-based chromatin activation, clarifying the spatial insulation function of CTCF, which needs to be explored further.

Discussion

The SCA-seq aimed to expand the traditional chromatin accessibility to high dimensional space by simultaneously resolving the chromatin accessibility and genome structure. Compare with 1D ATAC-seq, SCA-seq might more closely represent the relatively true structure of the native genome. With the SCA-seq, we found that the genome spatial contacts maintained the non-uniform chromatin accessibility, suggesting the complex genome regulation in 3D space. Further study with $CTCT^{-/-}$ indicated the insulating functions of CTCF on the spatial contacts. The first thing one needs to consider is efficiency of labeling. We used CTCF motifs to estimate labeling efficiency and the binomial test to correct the labeling accuracy at the single-molecule level. Then, relatively reliable molecular 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 the labeling efficiency may lead to deviation 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.

SCA-seq was created as a multi-omics tool to examine both chromosome conformation and chromatin accessibility. The second point that needs to be discussed is the 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. The precise open–open chromatin interactions were underdetermined. The alternative hypothesis is that the interactions locate outside the open chromatin (200 bp). Therefore, improvement of the resolution of chromosome conformation capture is needed to determine spatial accessibility interaction accurately.

Another important issue is the effect of CTCF knockout on chromatin conformation. In the previous publication, immediate removal of CTCF altered TAD boundaries [18]. However, the permanent depletion of CTCF in HEK293T cells had limited effects on genome architecture. We suspect that the changed chromatin condensability might affect the genome architecture and induce disorder in TAD boundaries. During long-term CTCF knockout, cells might compensate for the CTCF loss and stabilize genome architecture.

In this study, the permanent loss of CTCF activated spatially neighboring chromatin, which supported the insulating effect of CTCF. However, there were still over 30% of loci that could be activated without physical contact with CTCF motifs or peaks. We suspect that this phenomenon might be due to secondary effects of CTCF deficiency that promote cascade reactions and activation of spatially neighboring regions. However, it is possible that our definition of CTCF binding sites is not unambiguous as there may be more CTCF binding sites than we expected. In addition, although we have found that that loss of CTCF increased chromatin accessibility on spatial contacts, the reasons for such a clear causal relationship remain uncertain.

Overall, our results demonstrated that SCA-seq can resolve genome accessibility locations in the three-dimensional space, which paves 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] and CTCF knockout 293T cell line [abclonal] were each maintained in DMEM-high glucose [Thermo Fisher 11995065] supplemented with 10% fetal bovine serum (FBS)

[Thermo fisher 1009141]. The CTCF-/- cell line was purchased from ABclonal, and validated by western blot and qPCR.

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 mixure 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 reaction. 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].

SCA-seq pipeline.

We developed a reproducible bioinformatics pipeline to analyze the M.CviplI 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 fragment. 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 fragment on each concatemer is summarize by converting the filtered alignment to a fragment 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) [7]. The methylation call is then counted to each fragment in the fragment bed file to derive the methylated and unmethylated count of GpC and CpG for each fragment 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. The decomposed SCA-seq contact matrix was treated as a Hi-C contact matrix and analyzed by Hi-C software. 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 and TAD

insulation score. 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 [7]. 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.

Estimate the labeling efficiency in vitro

As previous research, the CTCF motif maintained the open 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 fragments with low CpG methylation were expected to maintain the open chromatin with CTCF binding. We hypothesized that the fragments with low CpG methylation (CpG ratio<0.25) and low chromatin accessibility (GpC ratio<0.1) were not efficiently labeled.

Filter the fragments by binomial test

The medium fragment length is 500bp, which is close to the general size of open chromatins. We first calculated the background level of the methyl-GpC(open) and non-methyl-GpC(close) probability on the fragments. We used the non-treated genomic DNAs as the background, and p(GpC background) were the average GpC frequency on fragments. Then we performed the binomial test (R basics) for each fragments 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 closed chromatin with the non-methyl GpC. The non-methyl-GpC frequency in M.CviPI treated spike-in close 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 fragments 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 fragments with adjusted p value less than 0.05. We determined the accessible fragments first, and then we further

determined the inaccessible fragments 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 fragment. However, the open regions in ATAC-seq were around 200bp (peak average size). If we used sliding windows (200bp windows, sliding 50bp) on each fragment, we may determine the precise accessible regions on the fragments 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 fragments 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.

Data availability

The data were stored at https://db.cngb.org/search/project/CNP0002862/.

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. NY and FR perform the lab experiments; YMX 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. Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell* **128**, 787-800 (2007).
2. 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).
3. Voong, L.N., Xi, L., Wang, J.-P. & Wang, X. Genome-wide Mapping of the Nucleosome Landscape by Micrococcal Nuclease and Chemical Mapping. *Trends in Genetics* **33**, 495-507 (2017).
4. 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).
5. Wang, Y. et al. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Res* (2019).
6. Abdulhay, N.J. et al. Massively multiplex single-molecule oligonucleosome footprinting. *Elife* **9** (2020).
7. Lee, I. et al. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nature Methods* **17**, 1191-1199 (2020).
8. Shipony, Z. et al. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat Methods* **17**, 319-327 (2020).
9. Chen, W. et al. Sequencing of methylase-accessible regions in integral circular extrachromosomal DNA reveals differences in chromatin structure. *Epigenetics & Chromatin* **14**, 40 (2021).
10. Weng, Z. et al. Long-range single-molecule mapping of chromatin modification in eukaryotes. *bioRxiv*, 2021.2007.2008.451578 (2021).
11. Lai, B. et al. Trac-looping measures genome structure and chromatin accessibility. *Nature Methods* **15**, 741-747 (2018).

12. 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).
13. Wei, X. et al. Multi-omics analysis of chromatin accessibility and interactions with transcriptome by HiCAR. *bioRxiv*, 2020.2011.2002.366062 (2020).
14. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).
15. McClelland, M. & Ivarie, R. Asymmetrical distribution of CpG in an 'average' mammalian gene. *Nucleic acids research* **10**, 7865-7877 (1982).
16. 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).
17. 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).
18. Ong, C.-T. & Corces, V.G. CTCF: an architectural protein bridging genome topology and function. *Nature Reviews Genetics* **15**, 234-246 (2014).
19. Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. *Nat Rev Genet* **20**, 437-455 (2019).
20. Özdemir, I. & Gambetta, M.C. The Role of Insulation in Patterning Gene Expression. *Genes (Basel)* **10** (2019).
21. Merkenschlager, M. & Nora, E.P. CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. *Annu Rev Genomics Hum Genet* **17**, 17-43 (2016).
22. Rowley, M.J. & Corces, V.G. Organizational principles of 3D genome architecture. *Nature Reviews Genetics* **19**, 789-800 (2018).
23. Kaushal, A. et al. CTCF loss has limited effects on global genome architecture in Drosophila despite critical regulatory functions. *Nat Commun* **12**, 1011 (2021).
24. Xu, B. et al. Acute depletion of CTCF rewires genome-wide chromatin accessibility. *Genome Biology* **22**, 244 (2021).

a

	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)	Paired-wise contact (2D)	Paired-wise contact (2D)	Not available (1D)	Not available (1D)	Paired-wise contact (1D)
Chromatin accessibility	Direct sequencing of accessible and inaccessible region	PCR enrichment of accessible region only	PCR enrichment of accessible region only	Direct sequencing of accessible and inaccessible region	PCR enrichment of accessible region 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 detecting spatial CpG regulation in the same concatemer	Not available	Not available	Simultaneous measuring CpG in the same singleton molecule	Not available	Not available

b

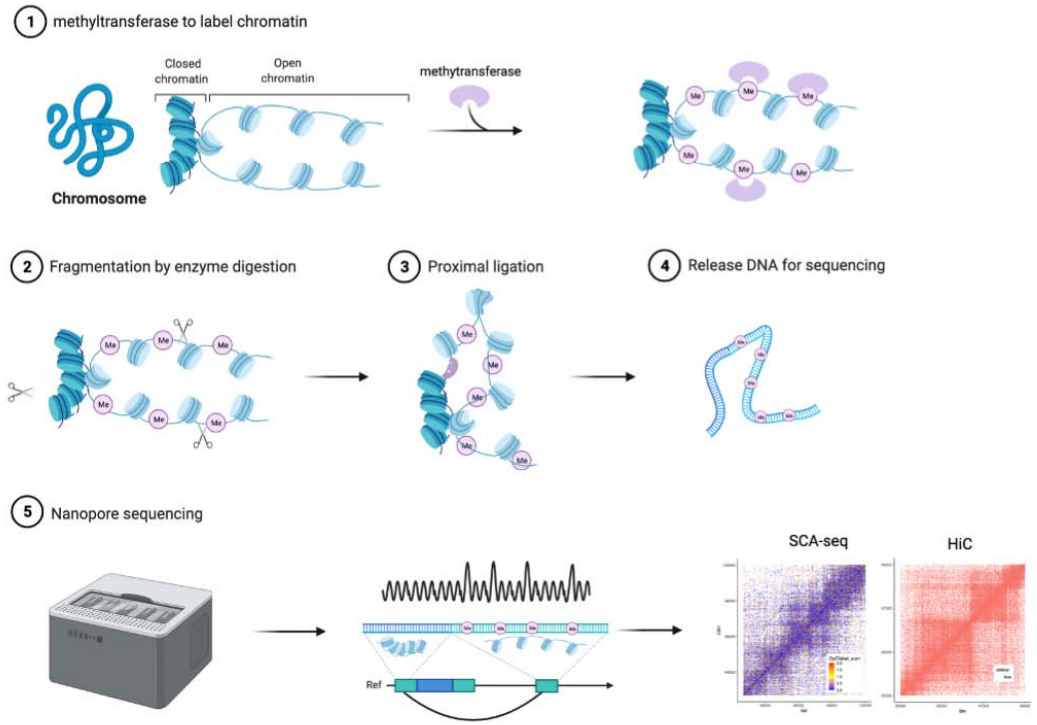


Fig1. The principle of the SCA-seq. We first compared the SCA-seq with other similar technologies (a). (1) After fixation, the chromatin accessibility could be labeled as artificial methylation by the methyltransferase (m6A or GpC, which are very few in native genome). (2-4) Then we selected the restriction enzymes to digest the labeled genome. The closed fragments (>2) stayed as the fragments cluster (**concatemer**) due to the fixation. Then the spatially close fragments in one concatemer could be proximately ligated, and formed the chimeric long fragments. (5) These long chimeric fragments

were sequenced by nanopore. And the data trained nanopolish model helped call the labeled artificial methylation (GpC or m6A), which carried the regional chromatin accessibility, and native CpG on the chimeric reads. Our algorithm analyzed the composition of the chimeric reads, indicating the genome locations of these composed fragments. The SCA-seq could generated the multiway HiC contact maps, which indicated the genome 3D conformation. Moreover, the SCA-seq could simultaneously produce the chromatin accessibility and native methylation information based on the contact maps.

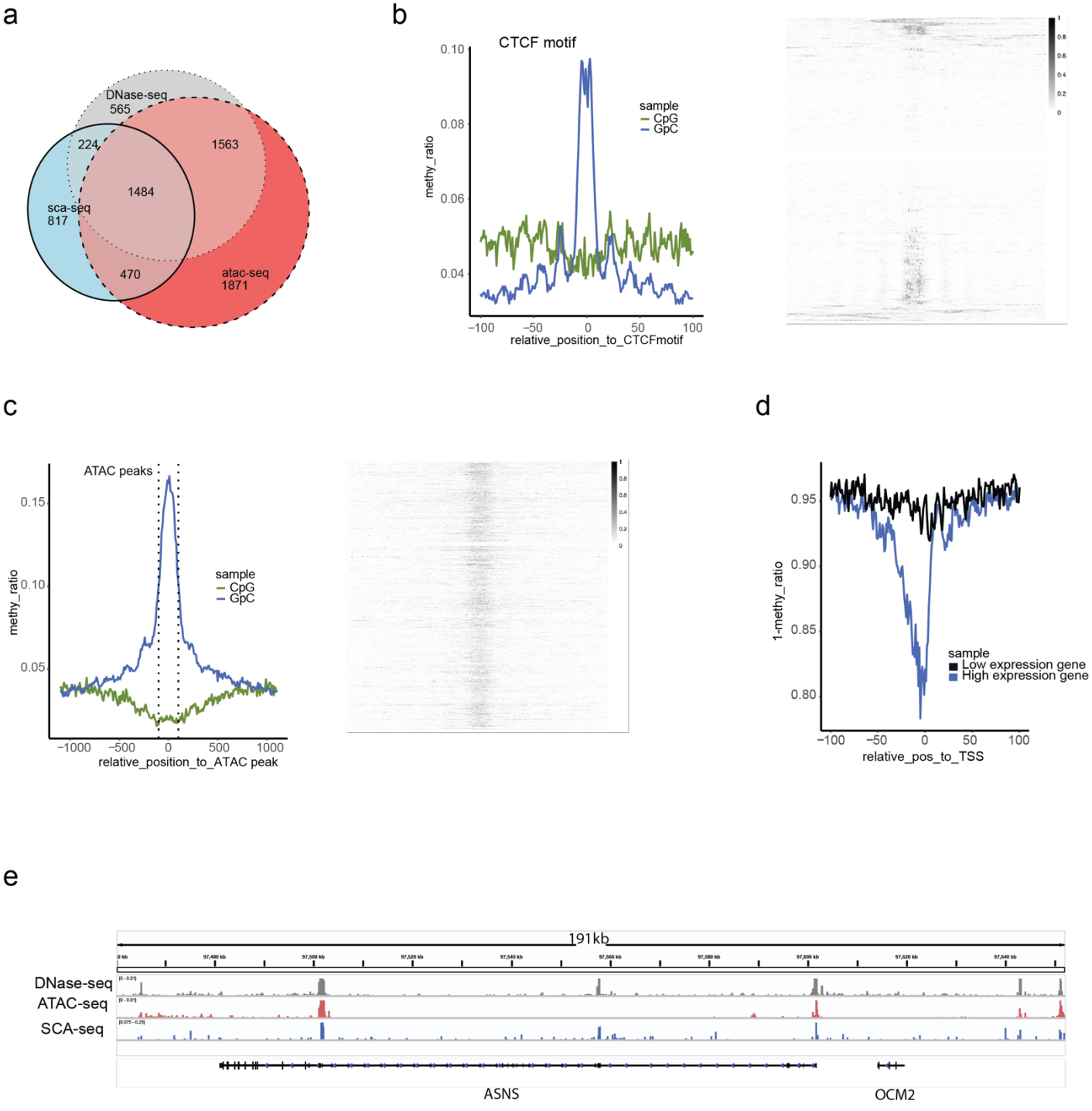


Fig2. The feasibility of SCA-seq to label the open chromatin. The peak calling algorithms were borrowed from nanoNOME-seq [1]. There are 72% peaks identified in SCA-seq, overlapping with either ATAC-seq or DNase-seq. We further plotted the CTCF motif centered plot (0 indicated the CTCF motif). Y-axis indicated the average methylation ratio (methylation ratio=methylated GpC/all GpC in 10bp bin) among all the molecules across CTCF motifs. The SCA-seq demonstrated the classic nucleosome depletion patterns (NDR) around the CTCF motif. The side panel demonstrated the single molecular resolution. Each line indicated one molecule across and the blue colors indicated the methylation ratio in 10bp bins. For similar methods, we centered the ATAC-seq peak regions (NDR regions) and normalized to 200bp (average peak size). The SCA-seq showed the similar chromatin accessible peaks in these regions. The side panel is the single molecular resolution demonstration. Furthermore, we classified the high expression genes (upper quantile) and low expression (lower quantile) genes by the expression ranks. The high expression genes showed the more drastic nucleosome depletion (y axis indicated 1-GpC methylation ratio) than the low expression genes in TSS (0). Panel e showed the raw data of DNase-seq, ATAC-seq, SCA-seq in genome browser (191kb span). DNase-seq and ATAC-seq showed the read counts in each genomic locus, and the SCA-seq showed the methylation site counts in each genomic locus.

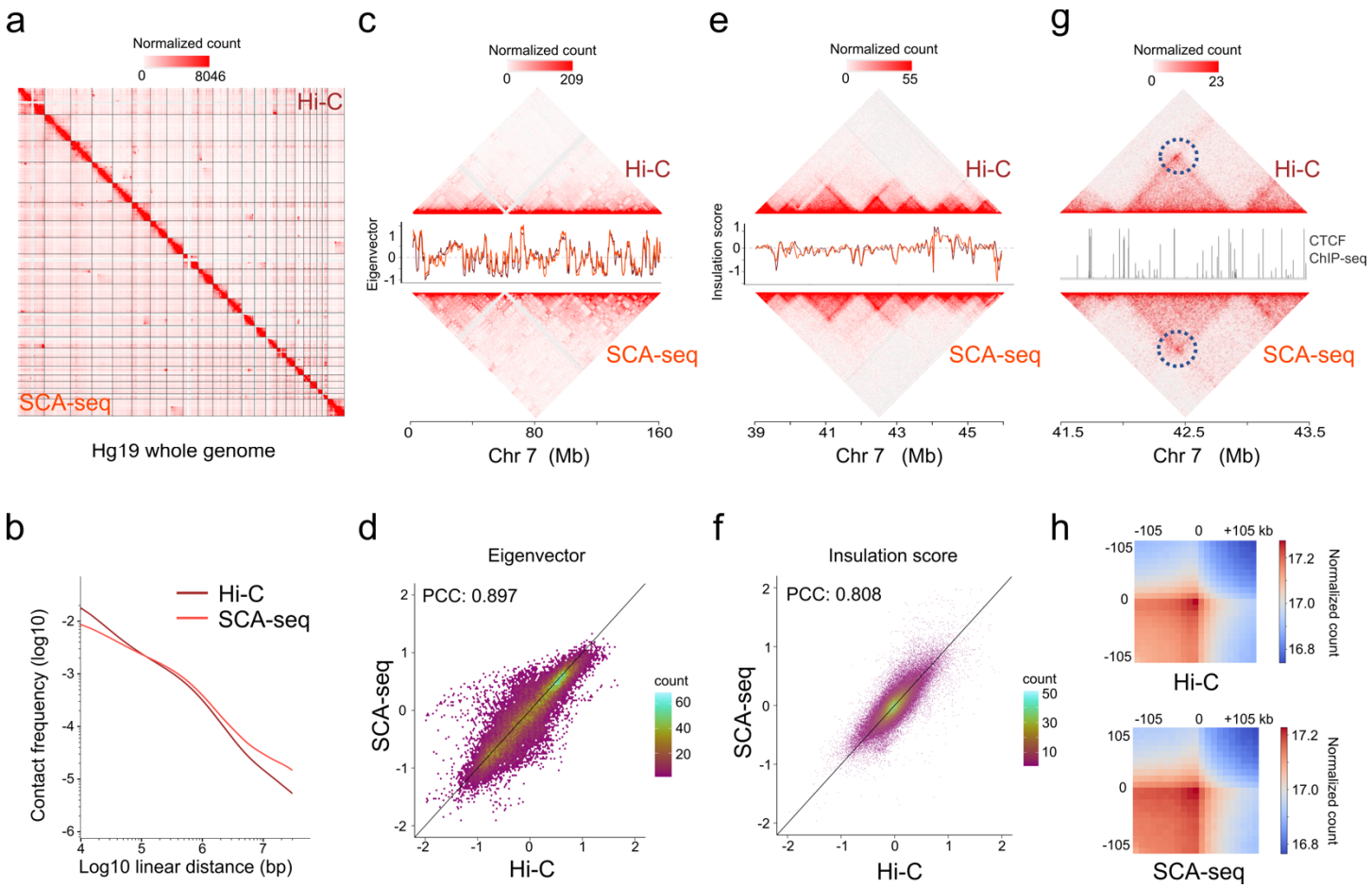


Fig3. (a)Comparison of 1.1 M SCA virtual pairwise contacts (lower triangle) and 1.8 M Hi-C. contacts (upper triangle) for chromosome 1-22 and X (hg19) in the cell line hek293T. (b) Contact frequency (y-axis) as a function of linear genomic distance (x-axis) was plotted across all hg19 chromosomes for SCA-seq (red) and Hi-C (brown). (c) and (e) Comparison of SCA and Hi-C (c) 250 kb and (e) 25 kb contact maps for chromosome 7. The corresponding (c) eigenvector and (e) insulation score indicates the visual consistency between SCA-seq (red) and Hi-C (brown) signal patterns. Color scale bar: log normalized read counts. (d) and (f) Scatterplots comparing (d) eigenvector and (f) insulation score between SCA and Hi-C. PCC: Pearson correlation coefficient. (g) Contact map showing an example of CTCF peaks at 10kb resolution with loop anchor signal indicated in the black circle. Color scale bar: log normalized read counts. The CTCF ChIP-seq tract was plotted in between. (h) Aggregate peak analysis (APA) showing the visual correspondence of enrichment pattern between Hi-C and SCA within 100kb of loop anchors at 10kb 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.

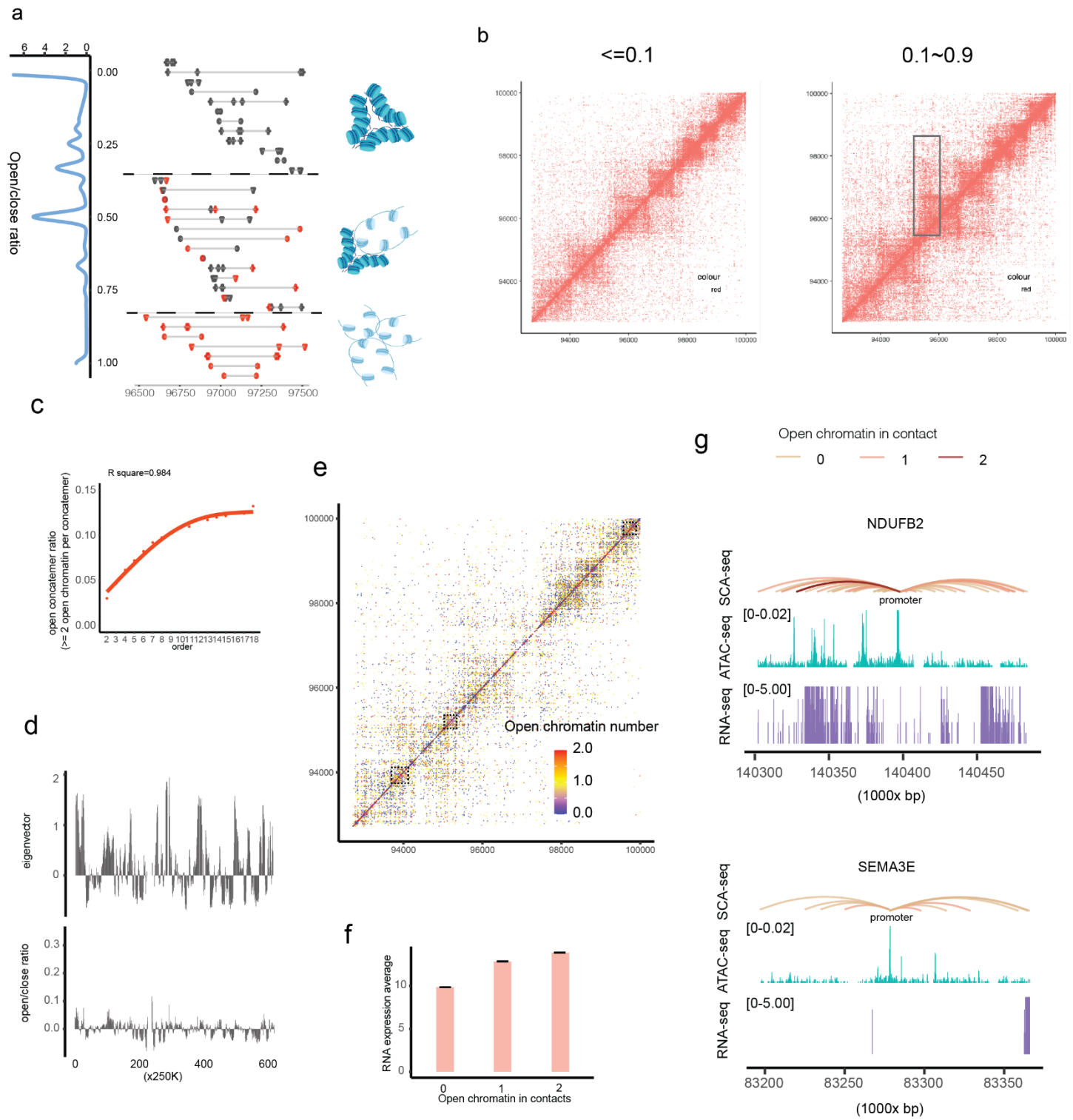


Fig4. The SCA-seq demonstrated the spatial chromatin accessibility. Each concatemer contained 2-10 genomic fragments. For each fragment, we determined it accessibility open or close. Then the open/close ratio is calculated as the $N_{open}/N_{open+close}$. As our observation, 29% concatemers (heterochromatin concatemer) contained >90% close chromatin fragments (open/close ratio<0.1). 62.2% concatemers (hybrid concatemer) contained 10%~90% close fragments (open/close ratio

0.1~0.9) and only 8.8% concatemers (euchromatin concatemer) contained >90% open chromatin fragments (open/close ratio>0.9). Then we plotted the contacted maps for the different types of concatemers in the genome region 92700000~100000000 with 1kb resolution (b). We subsampled 150000 contacts in both heterochromatin concatemer and hybrid concatemer for normalization. The euchromatin concatemers contained only 13000 contacts, which was too few to show any TAD structures (Not shown in figures). We found that the heterochromatin concatemers tend to gather inside the TAD and hybrid concatemers indicated more TAD boundary interactions and distal interactions (black box). We used the Juicer [2] to calculate the AB compartment on the chr7. In the overall statics, we also found open concatemers (open contacts >2 per concatemers) tend to positively correlated with the concatemer order (fragment number in each concatemer) . As above description, usually the high order concatemers indicate the long-distance interaction(c). Then we calculated the open/close average ratio on the corresponding compartment (Note: c showed the compartment open\close ratio-average open on chr7, showing the distinctions). We found the open/close ratio is highly correlated with AB compartment eigen values (p<2.6e-16, student t-test) (d). We further converted the contact maps to accessibility contact maps. The close chromatin pairwise contacts (close to close) were blue colors (0). The hybrid chromatin (open-close) pairwise contacts were yellow colors (1). The open chromatin pairwise contacts (open/open) were red colors(2). We plotted the enhancer/promoter interactions in the genome region 92700000~100000000. We found most of the regions were yellow colors, which is the open/close interaction. The open/open contacts gathered in the diagonal line and scattered surrounded the TAD (e). We then summarized the accessibility contacts relationship with the gene expression. We filtered the contacts carried promoters, then each contact could be linked to the gene expression under the corresponding promoters (gene upstream 2kb). We plotted the average gene expression with the different type accessibility contacts. The open/open chromatin contacts could significantly enhance the gene expression (p<2.6e-16, student t-test)(f). To further explore the impact of the open/open contact, we selected two genes with the similar accessibility on the promoter (2.1% fragments opens, promoter

centered plot). NDUFB2 with more open/open contacts carried much higher gene expression (528 RPKM) than SEMA3E (0 RPKM), which carried few open chromatin contacts(g).

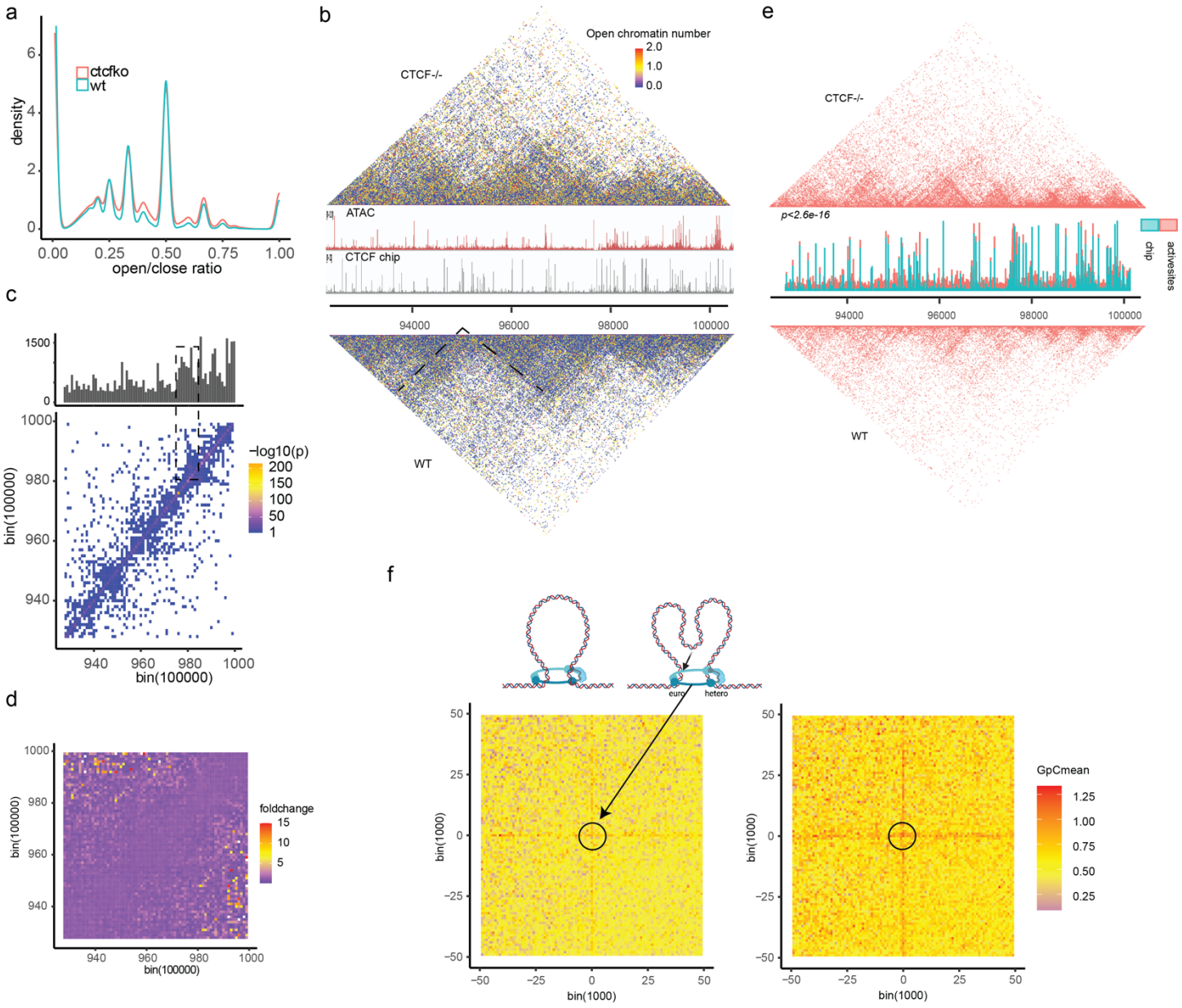
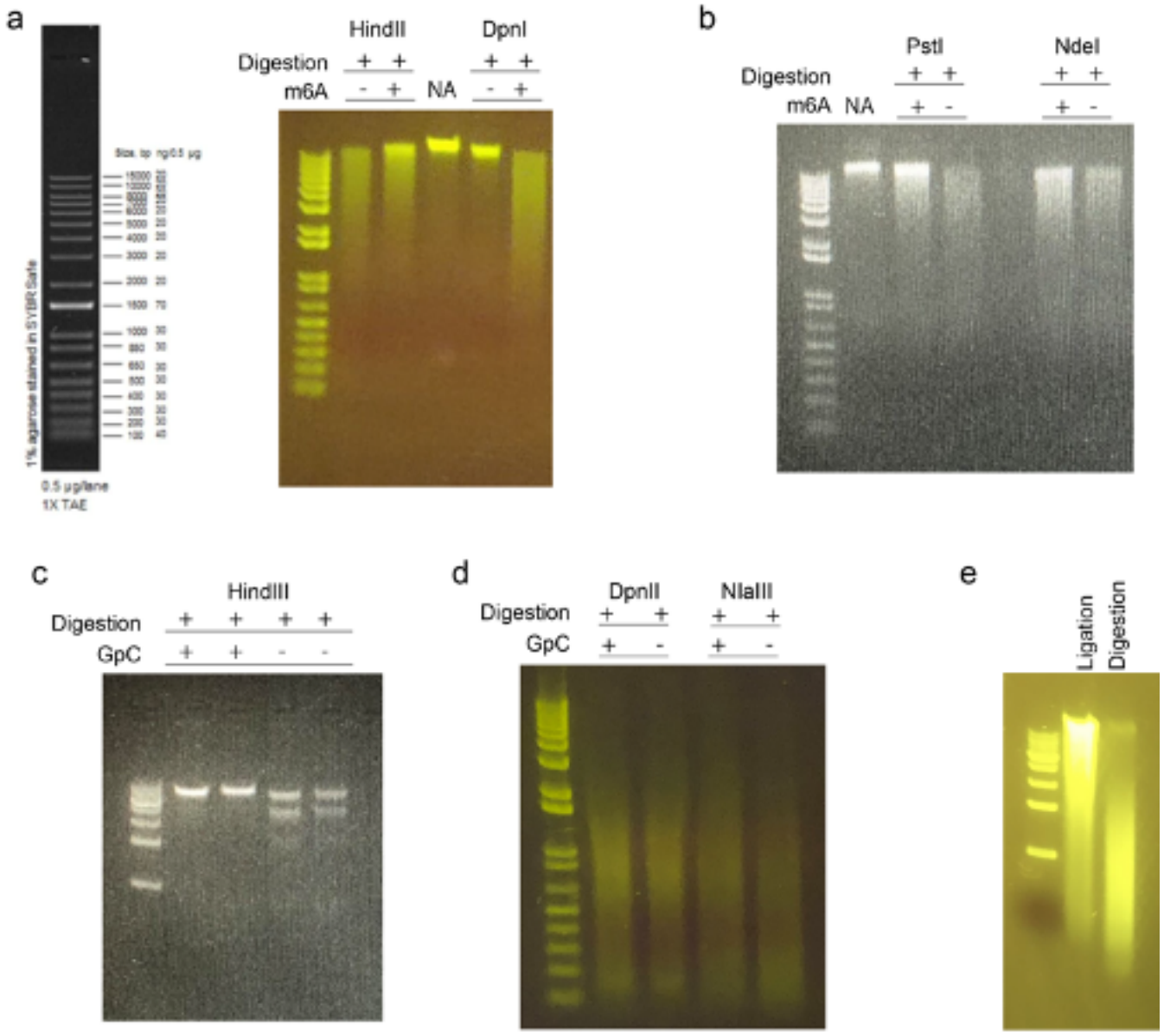


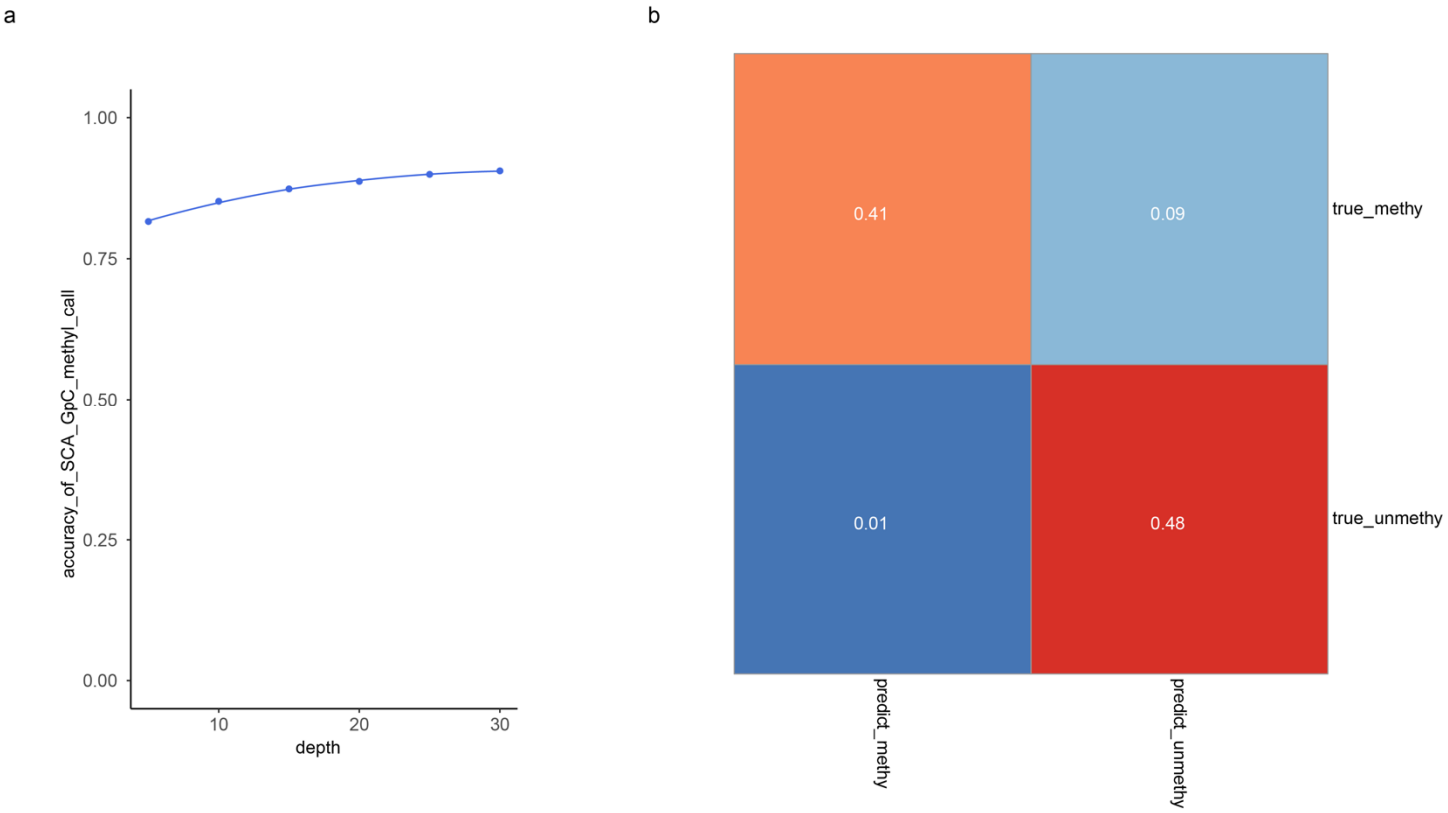
Fig5. The genome was activated by contacts in CTCF-/- HET293T cells. In the CTCF-/- cells, the concatemers contained significantly more open chromatin than the WT (*t.test p<2.6e-16*) (a). We further converted the contact maps to accessibility contact maps (The close chromatin contacts (close to close~0 blue; the hybrid chromatin (open-close) contacts were 1 yellow; the open chromatin contacts (open/open) were 2 red.). As observed in WT, the hybrid chromatin contacts (yellow line) were distributed in lines or chessboard patterns, which were overlapped with CTCF peaks or ATAC peaks (b). In the contrast, the CTCF-/- significantly enhanced the overall proximal accessibility in the contact maps (*Wilcoxon p<2.6e-16*) (b). To further study which regions were affected, we performed the regional Wilcoxon test with the 100000bp square on the accessibility contact maps and corrected

by the Benjamini & Hochberg method. The significant changed region with adjusted p values (<0.1) were plots on the contact map(c). The significantly changed areas were distributed with the CTCF signals (*Pearson p<0.05*) (c). In the similar concept, we calculated the fold change in each region. The fold change in the TAD-outside regions were much larger than in the TAD-inside regions (student *t-test*, *p*<2.6e-16) (d). The CTCF bind sites maintained open chromatin (Fig2b). Most genome regions contacted with CTCF binding locus still maintained their original close or open chromatin. These contacts showed the yellow or red colors, which formed in lines in WT (b). Theoretically, the CTCF loss largely activated the accessibility contact maps and transformed the contacted locus to open chromatins (open-open contacts). Therefore, we selected the open-open contacts in the CTCF-/-. Obviously, the open-open contacts were significantly more in CTCF-/- than in WT. The open-open contacts were distributed in the line patterns. In the further studies, we sum up the open-open contacts based on the genome bins in CTCF-/- (number of active sites). The number of active sites were highly correlated with the CTCF chip signals (*Pearson p<2.6e-16*), suggesting the most increased open-open contacts interacted with the CTCF binding sites. We used the similar APA algorithm to search the CTCF loops by the CTCF peaks. Then we overlayed the 196 loops together to increase the accessibility signals. The x-axis bin 0 indicated the CTCF peak, and the y-axis bin 0 indicated the partner CTCF peak, which formed the CTCF loops. We plotted the upstream and downstream 50000bp. The color indicated the average contact accessibility from the overlayed 196 loops. As observed in WT, some DNAs contacted accessibilities might be activated by the interaction with CTCF loops, forming the cross pattern. The middle accessibility (CTCF loops location) of the cross is weak. Some CTCF loops may span the active and inactive genome, and the heterogenous accessibility were insulated by CTCF. In the CTCF-/-, most DNAs contacted accessibilities significantly activated by the interaction with CTCF/cohensin loops (cross pattern turn red), especially the middle of the cross pattern.



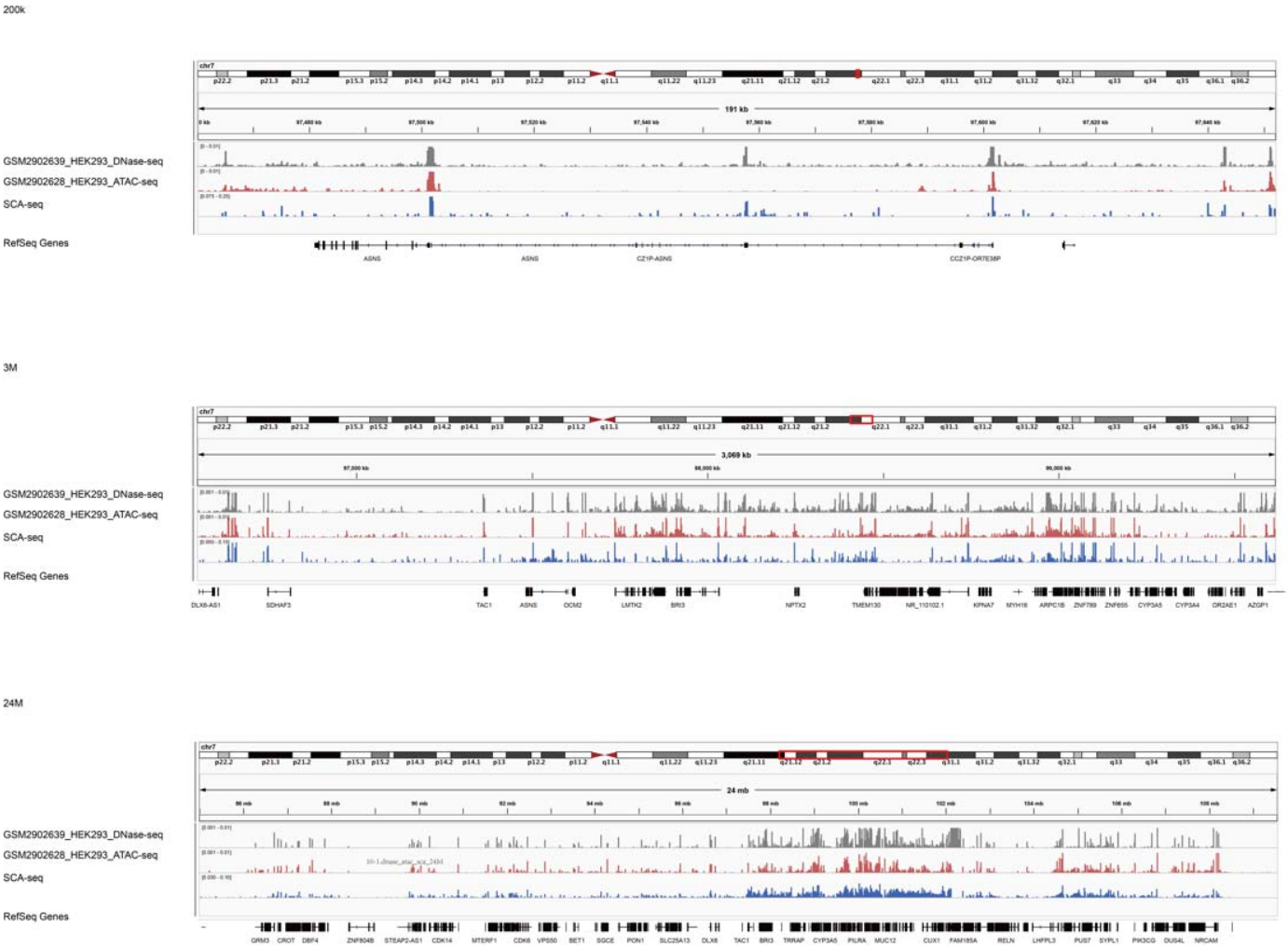
Sfig1. Selecting the compatible restriction enzymes. In the SCA-seq, we used the methyltransferase (EcoGII or M.CviPI) to label the open chromatin. Then we used the restriction enzyme to digest genome and prepare for the next step ligation. The EcoGII methyltransferase transferred the methyl group to the 6' carbon in adenosine (m6A modification). Due to the high density of the adenosines on genome (~25%), the abundant artificial m6A may impair the restriction enzymes. We used the EcoGII to treat the HEK293T genomic DNAs (m6A +), and leave the control (m6a -) untreated. We tested HindIII (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 the m6A dependent restriction enzymes, and only digested the genomic locus with m6A modification (a DpnI). NA was the

genomic DNA control. By the similar process, we tested the HindIII (c for lambda DNA), DpnII and NIaIII (d for HEK293T genomic DNAs) on the GpC methylated DNAs, which were treated by methyltransferase M.CviPI. DpnII and NIaIII were not significantly inhibited by GpC methylation. In the next step, we examined if the digested products could be ligated by T4 DNA ligase (e).

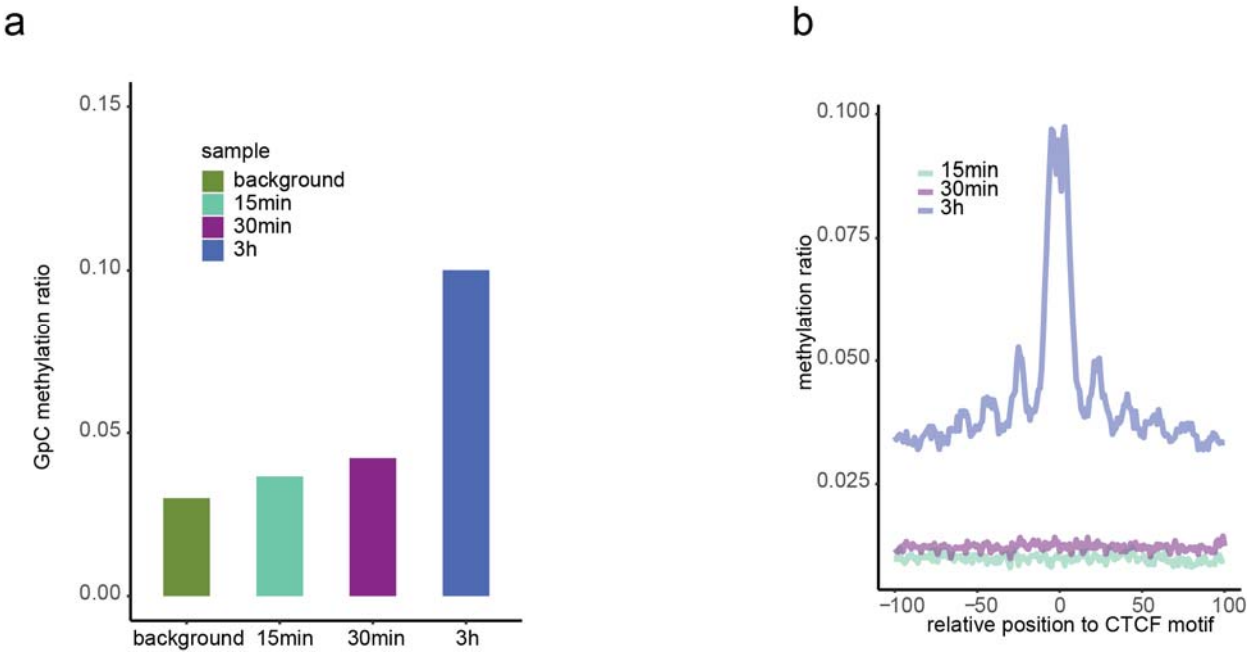


Sfig2. The feasibility of the GpC calling. (a)Similarity between the gold standard WGBS GpC methylation calls and SCA-seq GpC methylation calls at varies sequencing depth. GpC sites in 200bp bin of hg19 genome are classified as methylated and unmethylated by binomial test to reject the null hypothesis of unlabeled lambda DNA background GpC methylation signal. If a genomic bin is classified as methylated by both WGBS and SCA-seq, this SCA-seq methylation call is considered true positive. If a genomic bin is classified as unmethylated by both WGBS and SCA-seq, this SCA-seq methylation call is considered true negative. Then the accuracy of SCA-seq methylation calls is calculated at a gradient of maximum depth of genomic regions. (b) Confusion matrix of GpC

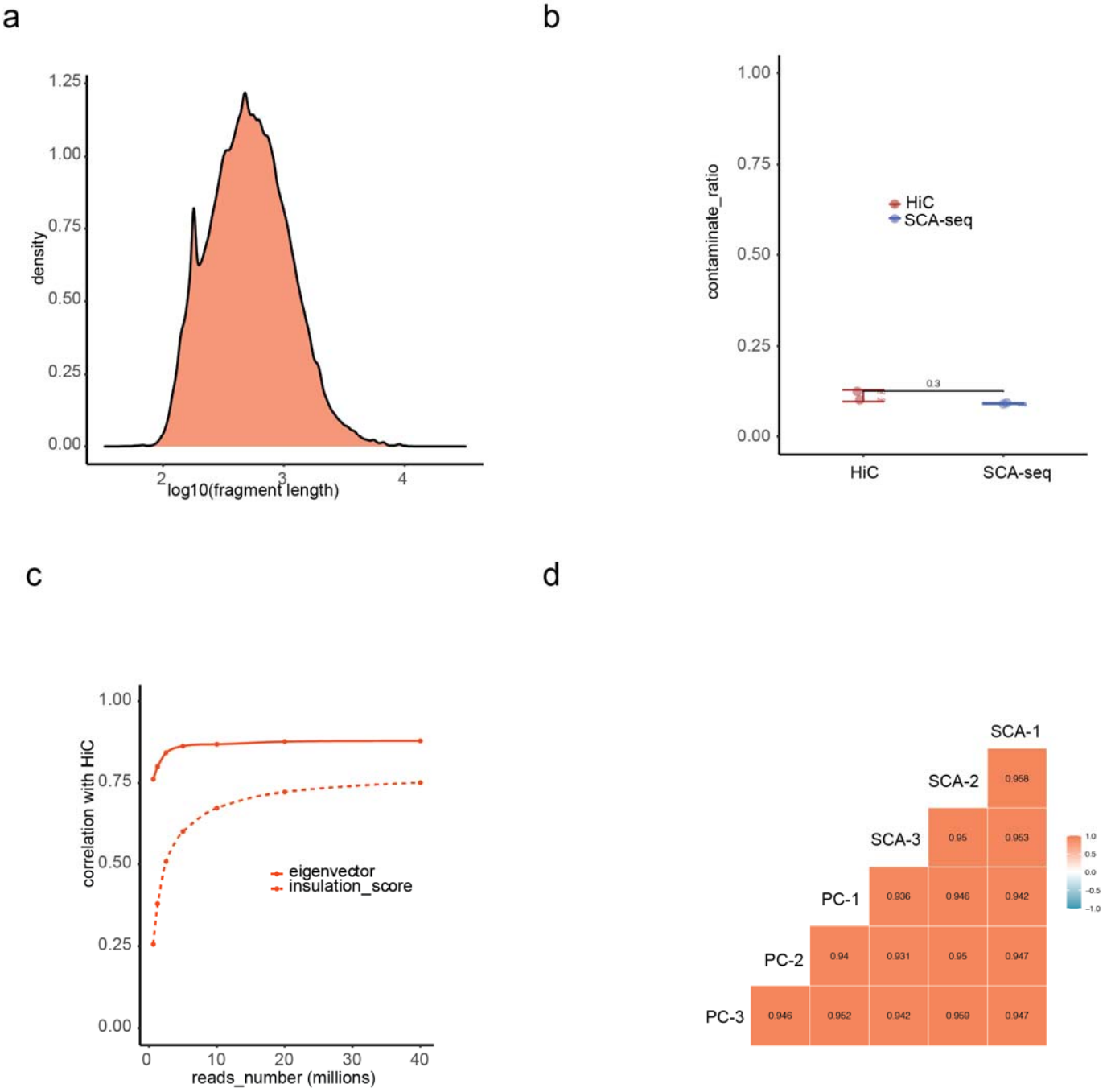
methylation calling. Subsets of 100,000 sites in unmethylated poreC hek293T samples and methylated lambda DNA samples were methylation called by default threshold of nanopolish cpggpc model. The results are classified into 4 categories, specifically, false positive and true negative from unmethylated poreC run and true positive and false negative from GpC labed lambda DNA sample.



Sfig3. The chromatin accessibility in various resolution. The panels showed the raw data of DNase-seq, ATAC-seq, SCA-seq in genome browser. DNase-seq and ATAC-seq showed the read counts in each genomic locus, and the SCA-seq showed the methylation site counts in each genomic locus.

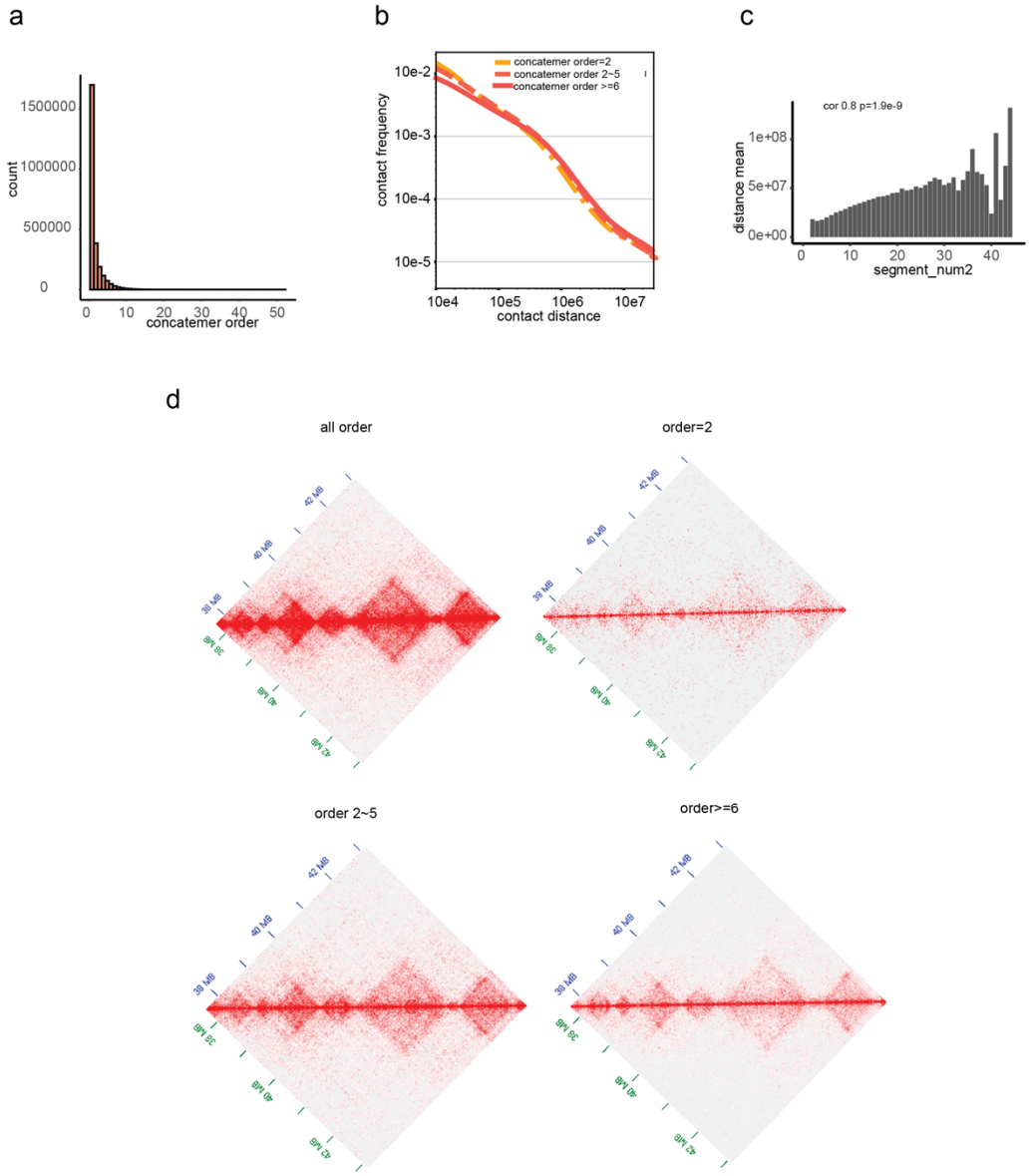


Sfig4. The labeling efficiency and signal strength in treatment dose. We tested three different labeling times(15 min 2 times added enzyme, 30 min 2 times added enzyme, 3 hours 3 times added enzyme). As our observation, 3h treatment showed the highest GpC labeling efficiency (a) and the labeled GpC signal accurately showed the CTCF motif nucleosome pattern (b). The 3 hours treatments were also similar to the nanoNOME-seq tissue sample preparation[1].



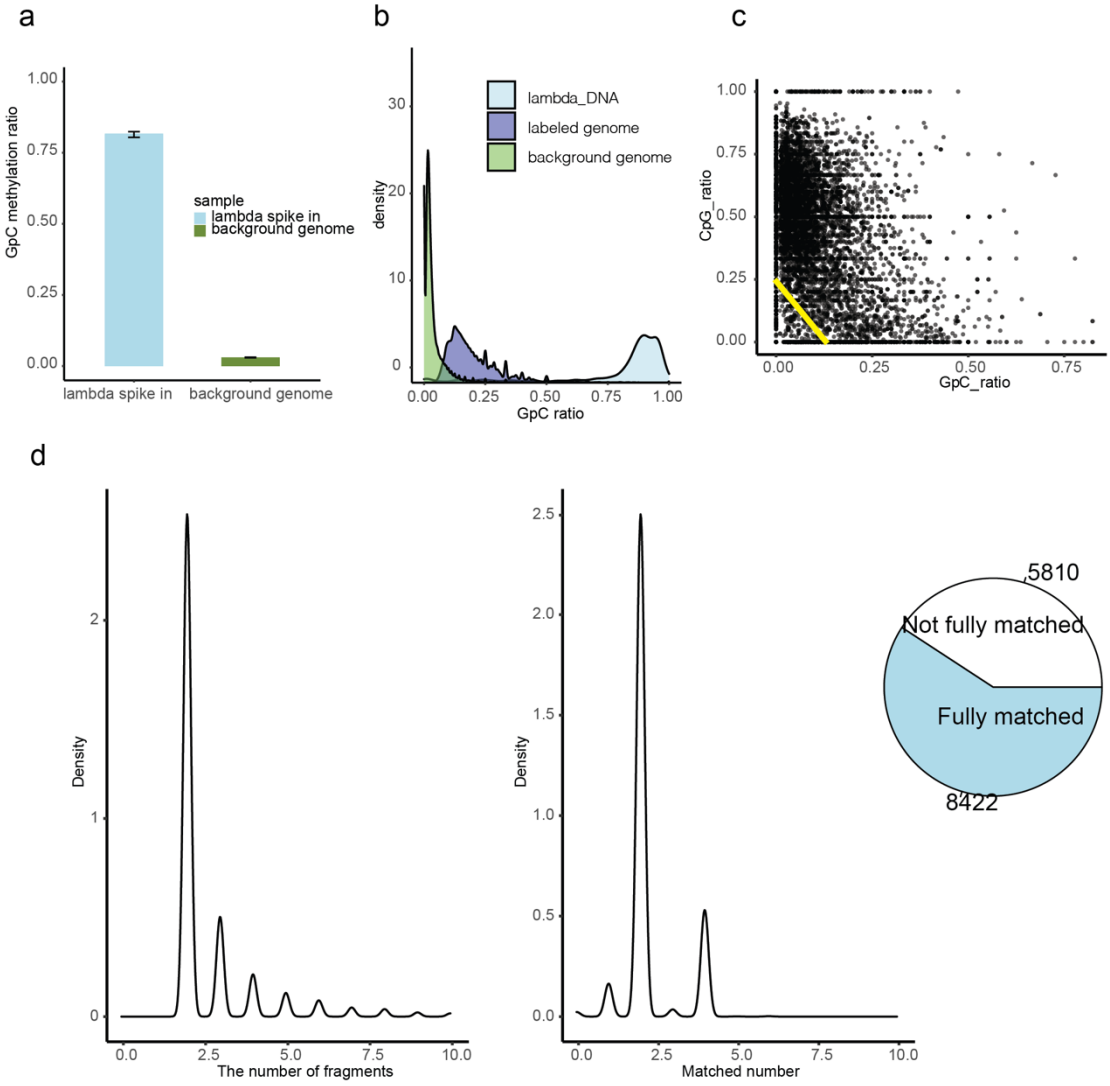
Sfig5. The quality statics of SCA-seq for resolving the genome structure. The medium length of fragments in concatemers were around 700bp (a). We further calculated the false positive contacts caused by the ligations (contamination). We used the contacts between mitochondrial genome and nucleic genome/total contacts to calculate the contaminations. There were 9% contamination in SCA-seq, comparing with the Hi-C 12% contamination. We then subsampling the reads to test the sequencing threshold to calculate the eigenvector (AB compartment) and insulation score (TAD). We need the over 5 millions and 30 millions reads to accurately produce the Hi-C similar results in AB

compartment and TAD. We then tested the correlation (0.93~0.94) between SCA-seq replicates and pore-C replicates [3]. SCA-seq demonstrated the good consistency between samples.



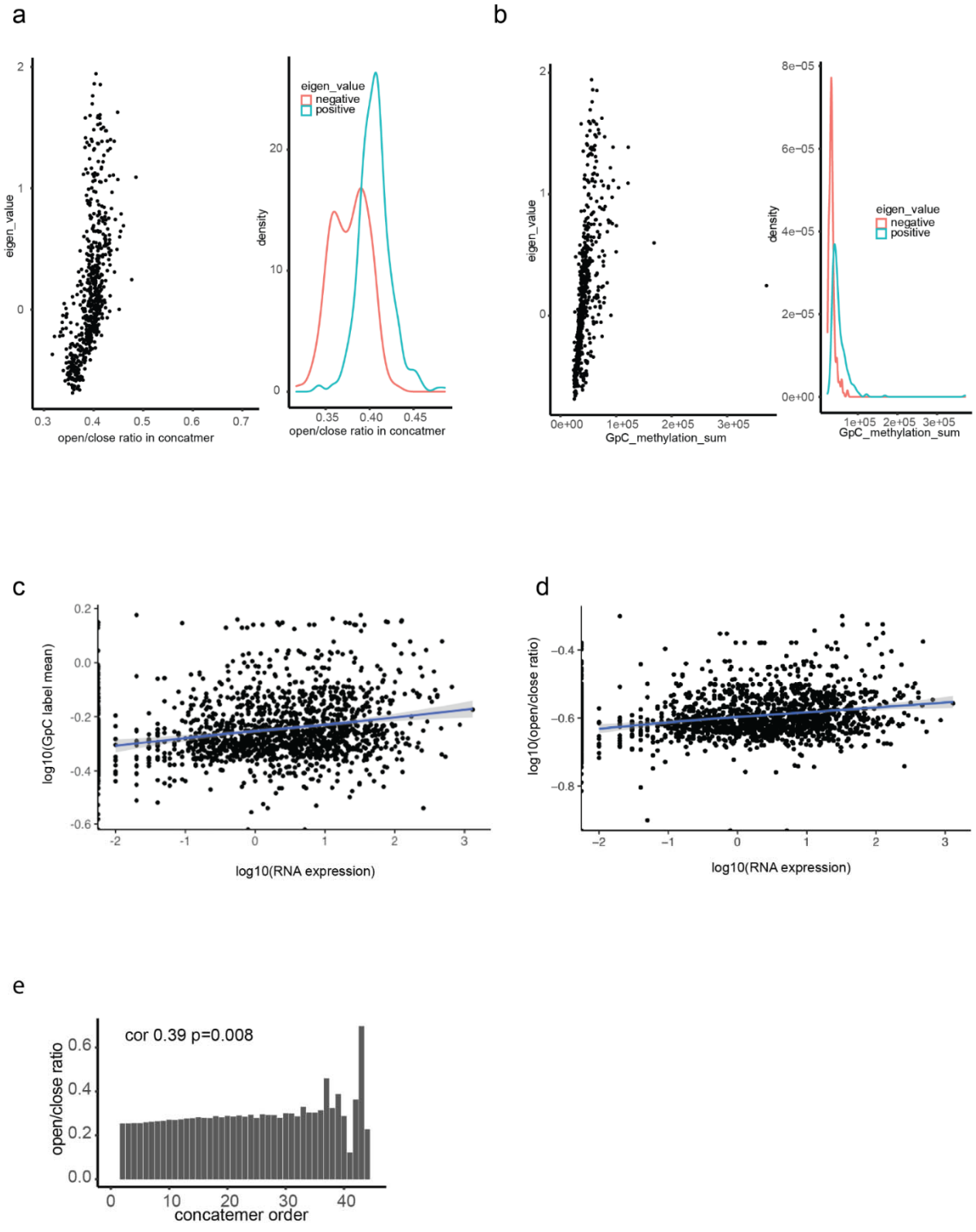
Sfig6. The high order contacts resolved the long-distance genome interaction. 65% concatemers contain 1 fragment without any genome contacts. 14.7% concatemers contain 2 fragments and 14.5% contained the 3-5 fragments. 5.4% concatemers have >=6 fragments (a). We plotted the contact frequency with the contact distance in different concatemer orders. The high order concatemers have more distant contacts than the low order concatemers (b). The maximum contact distance in each concatemer also positively correlated with the concatemer order (c). We converted the concatemers

with different orders to the contact maps. The high order concatemers tend to maintain the TAD structures and indicated long distance interaction (d).



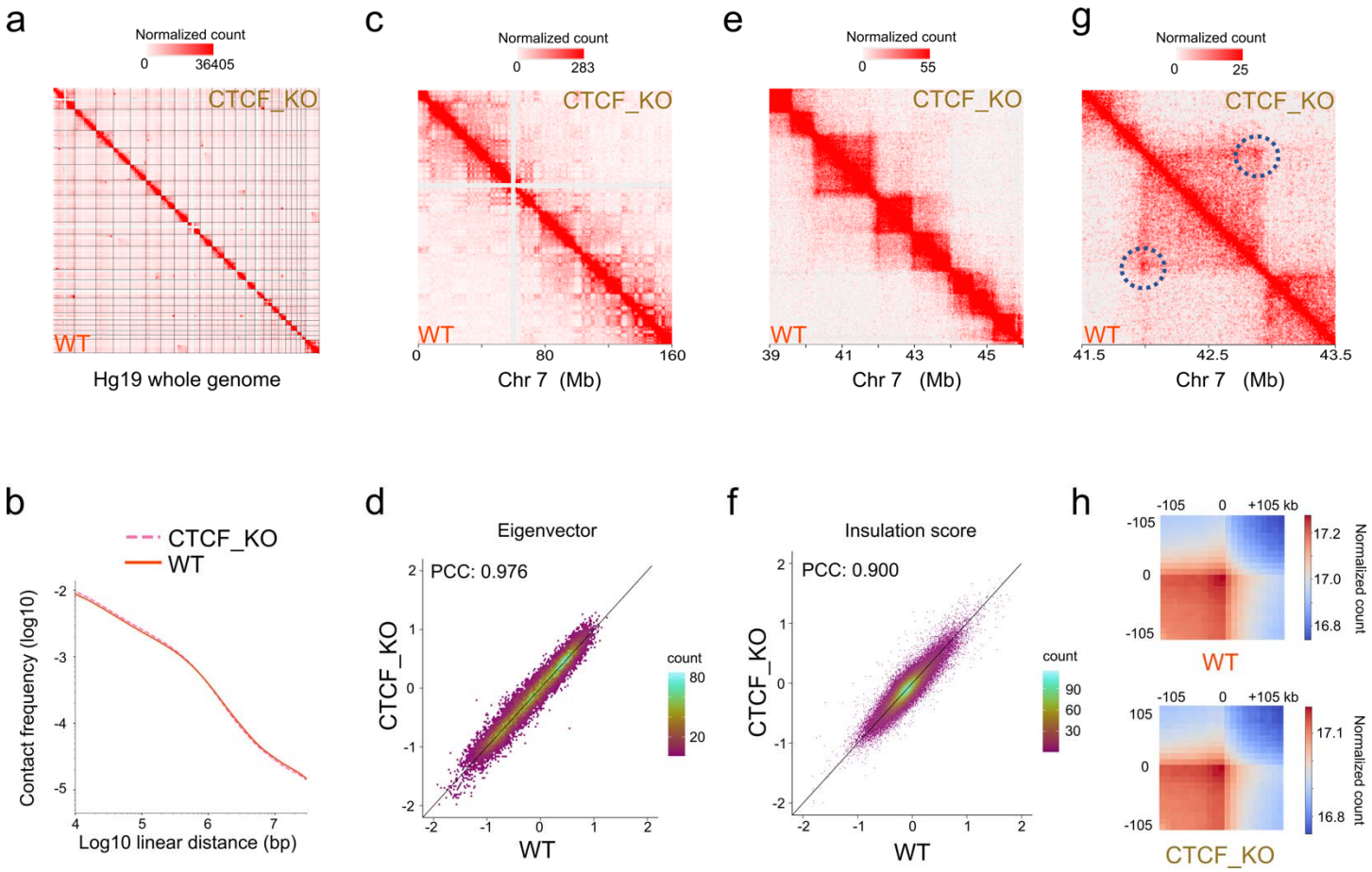
Sfig7. The GpC labeling efficiency in vitro and in vivo. We used the lambda DNA as the spike-in DNA control to monitor the M.CviPI efficiency. The genomic DNAs without labeling were used as the

background control to calculate the background level of GpC. The labeled genomic DNAs were processed as SCA-seq. We plotted the density plot of GpC ratios in each segment (methylated GpC/total GpC in each segments) for background genomic DNA, treated genomic DNA, spike-in control. The background level of GpC ratio in each segment were around 0.018 (a,b). The spike-in lambda DNA showed the M.CviPI labeling efficiency is around 0.88 in vitro (a,b). We then selected the CTCF motifs as the in vivo control to calculate the in vivo labeling efficiency(c). We plotted the CpG methylation ratio (methylated CpG/total CpG) with the GpC methylation ratio in CTCF motifs. Previous publications showed the pre-existing methylation can antagonize CTCF binding in CTCF motifs [4-6]. Therefore, we could assume that the CTCF motifs with low methylation (CpG methylation<0.25) maintained as the open chromatins with the CTCF binding. Then 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 the CTCF to maintain the open chromatin. Therefore, the in vivo labeling efficiency only could be roughly estimated. We then used the nanoNOME-seq to cross validate the labeling efficiency of SCA-seq(d). Both nanoNOME-seq and SCA-seq were single molecular assays. nanoNOME-seq could observe the chromatin accessibility in 2~20kb single DNA molecule. Literately, the accessibility of concatemers in SCA-seq could find the dedicated DNAs in nanoNOME-seq with the same accessibility on corresponding genome loci. The left panel (d) showed the number of fragments in the concatemers. 68% labeled concatemers could find the dedicated DNAs in nanoNOME-seq (c right panel). The most matched concatemers contained only 2 fragments, representing the short distance interaction. The concatemers with multiple fragments, representing the long-distance interaction, may be out of the detection range with nanoNOME-seq.

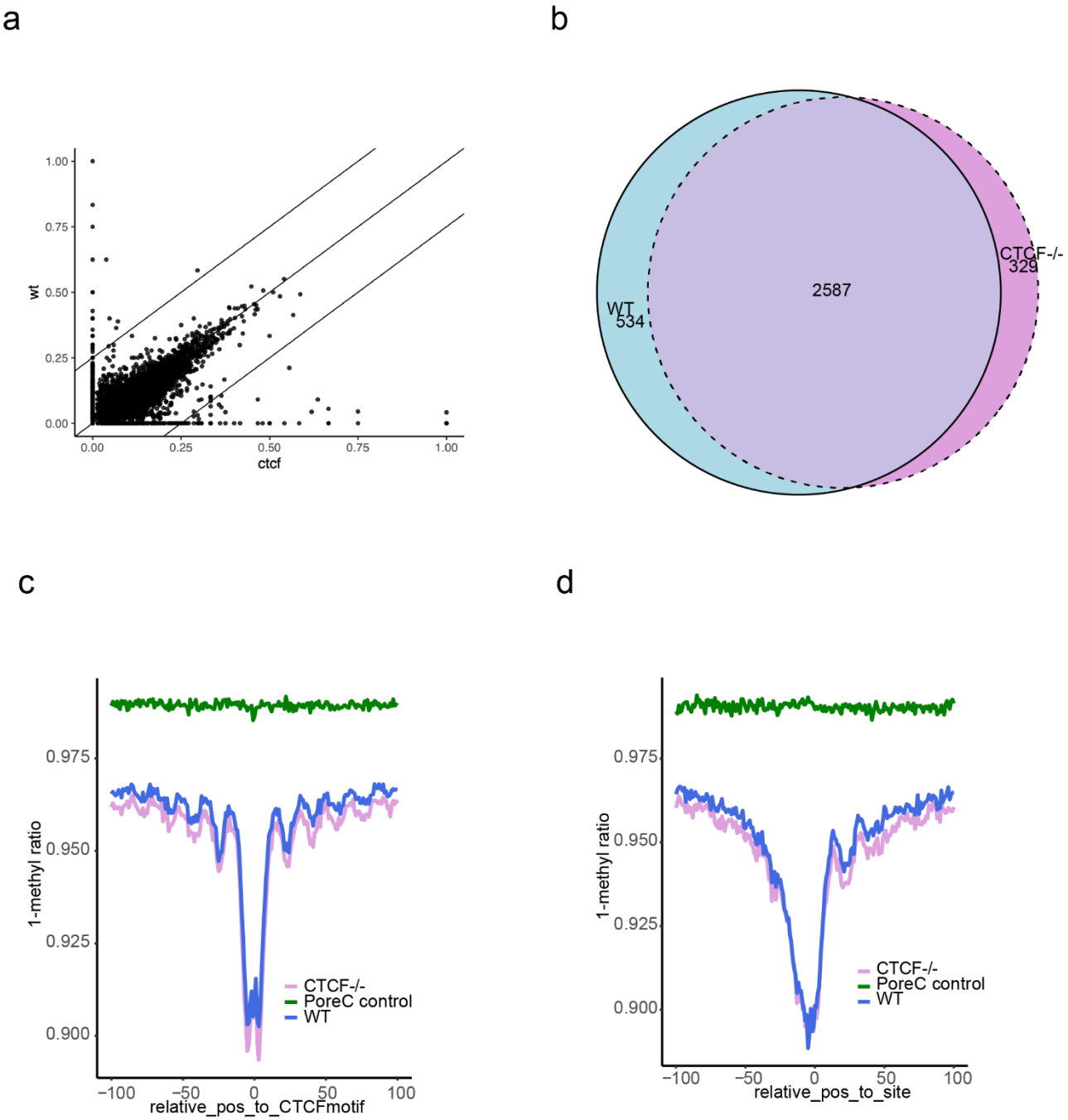


Sfig8. The correlation between compartmental eigen value, RNA expression and spatial accessibility. (a)We plotted the compartmental eigen values with the average concatemer open/close ratio in the compartments. The average concatemer open/close ratio is higher correlated with the compartmental eigen value (Pearson correlation *p*<2.6e-16). The average concatemer open/close ratio was

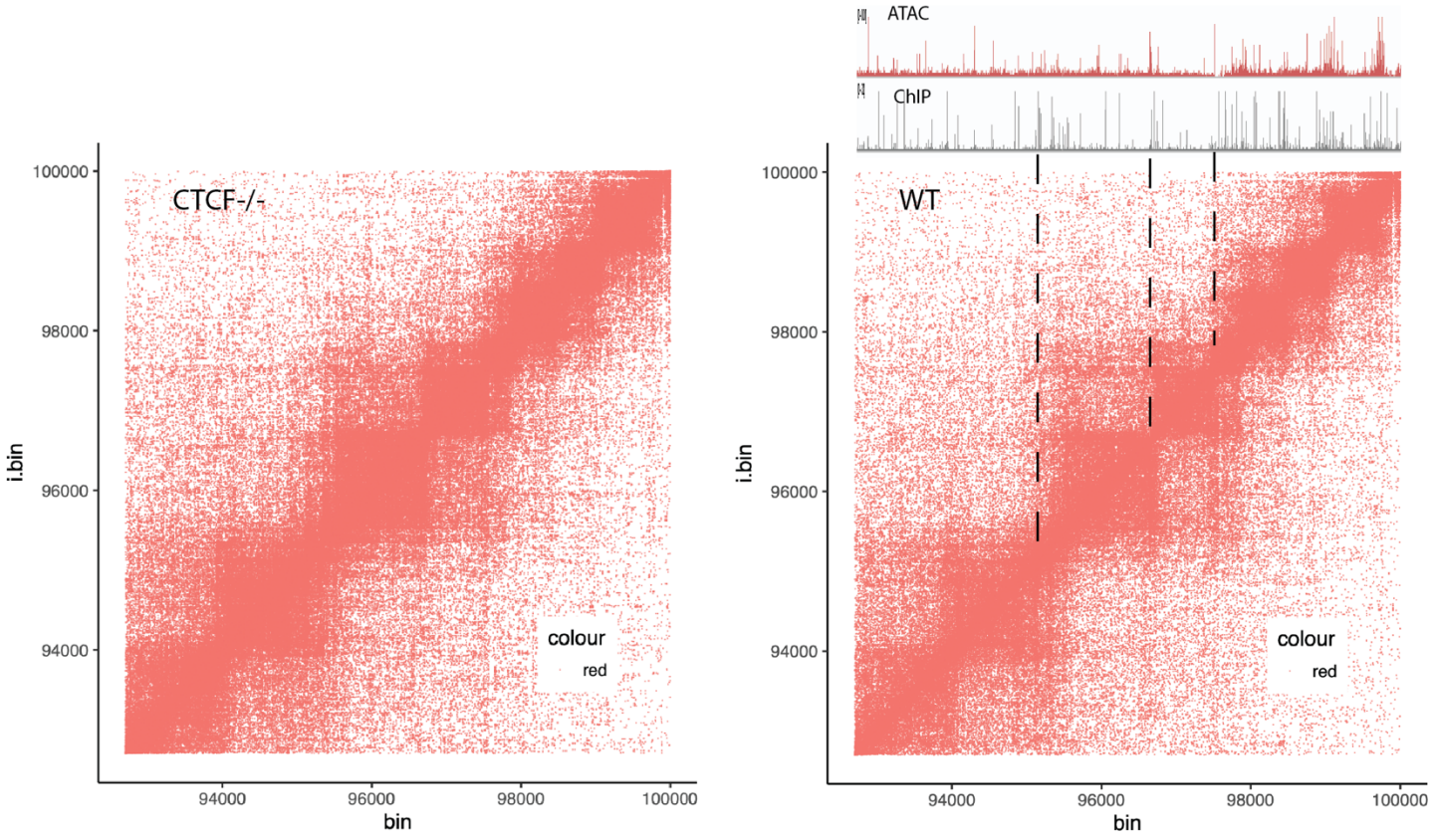
significantly higher in A compartments (positive eigen value) than in the B compartments (negative eigen value) (p<2.6e-16, student *t-test*). (b)In the similar way, we plotted the compartmental eigen values with the labeled GpC methylations. The labeled GpC methylations were high correlated with the compartmental eigen value (Pearson correlation p<2.6e-16). The labeled GpC methylations were significantly more abundant in A compartments (positive eigen value) than in the B compartments (negative eigen value) (p<2.6e-16, student *t-test*). (c) For enhancer/promoter accessibility contacts in Fig4d, we summarize the mean contactable accessibility on the promoters (close/close 0, close/open 1, open/open 2). The mean contactable accessibilities on promoters were highly correlated with their gene expression (Pearson correlation p<2.6e-16). (d)Similarly, we summarized the mean open/close ratio on the concatemers with the promoter regions. Then average concatemer open/close ratios on promoters were significantly correlated with gene expression (Pearson correlation p<2.6e-16).(e) The open/close ratios of concatemers were positively correlated with the concatemer order (Pearson correlation $p=0.0008$). In the Sfig6bc, we found the high order concatemers indicated the long-distance interactions.



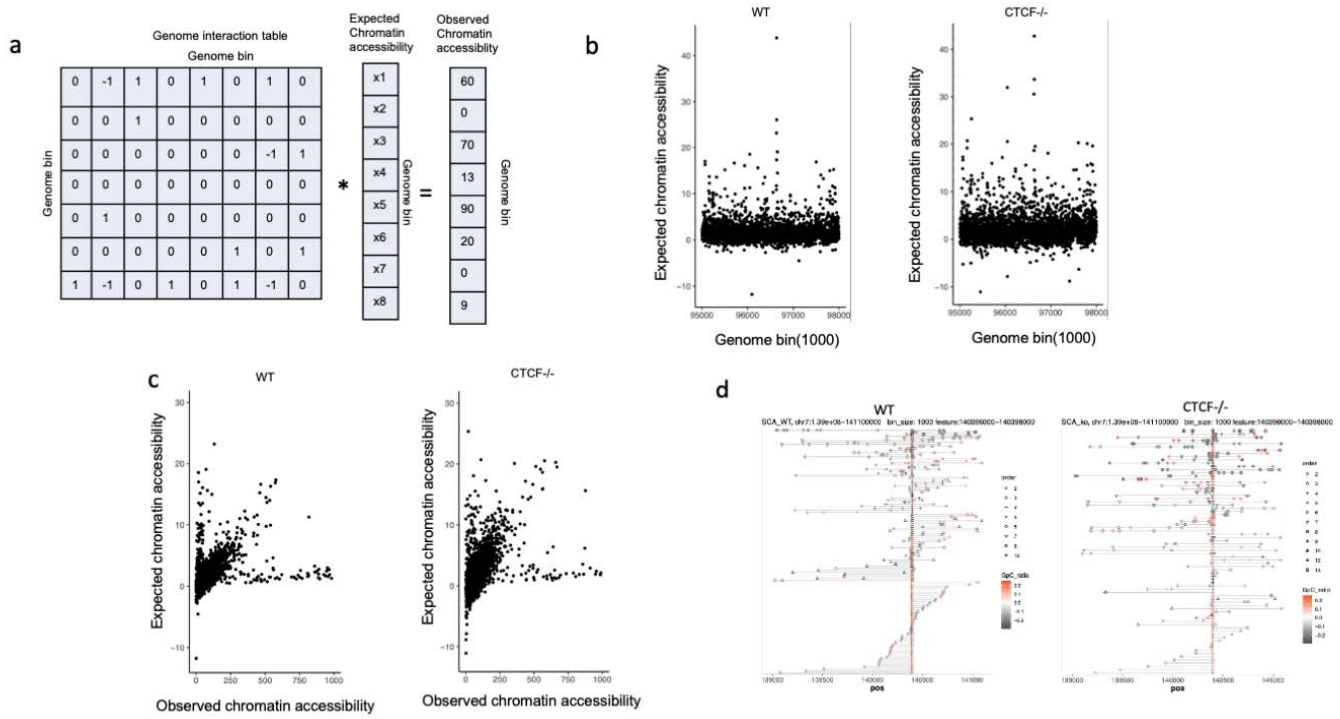
Sfig9. Compare the genome architecture between WT and CTCF-/-. (a)Comparison of 1.11 M virtual pairwise contacts (lower triangle) of SCA-seq hek293T wild type (WT) and 1.16 M virtual pairwise contacts (upper triangle) of SCA-seq hek293T CTCF knockout (CTCF_KO) for chromosome 1-22 and X (hg19). (b) Contact frequency (y-axis) as a function of linear genomic distance (x-axis) was plotted across all hg19 chromosomes for WT (solid line) and CTCF_KO (dashed line). (c) and (e) Comparison of WT and CTCF_KO (c) 250 kb and (e) 25 kb contact maps for chromosome 7. Color key: log normalized read counts. (d) and (f) Scatterplots comparing (d) eigenvector and (f) insulation score between WT and CTCF_KO. PCC: Pearson correlation coefficient. (g) Contact map showing an example of CTCF peaks at 10kb resolution with loop anchor signal indicated in the black circle. Color key: log normalized read counts. (h) Aggregate peak analysis (APA) showing the visual correspondence of enrichment pattern between Hi-C and SCA within 100kb of loop anchors at 10kb resolution. Color key: sum of contacts detected across the entire loop sets at CTCF ChIP-seq peaks in a coordinate system centered around each loop anchor.



Sfig10. Compare the chromatin accessibility between WT and CTCF-/-. We segmented the genome into 1000bp bin and calculated the methylation ratio in the genome bins. The lines indicated the two-fold change range. Most chromatin accessibility in most regions were not significantly changed. With accessibility peak calling algorithm, 534 accessibility peaks were downregulated and 329 accessibility peaks were upregulated in the CTCF-/-. The CTCF-/- showed the similar nucleosome distribution around the CTCF (c) motif and TSS(d).



Sfig11. The distribution of the open/half-open chromatin contacts. The SCA-seq data were converted to the 2-way contacts. The close-close chromatin contacts were defined as 0; the close-open chromatin contacts were defined as 1; the open-open chromatin contacts were defined as 2. We filtered the open/half open contacts with accessibility >0. In WT, the open contacts were distributed as lines/chessboard, which overlapped with the CTCF ChIP signal and ATAC peaks. In CTCF-/-, the open contact lines were blur and most open contacts were distributed in the cores.



Sfig12. The expected chromatin accessibility. The chromatin accessibility could be activated by the contacts with spatially close genome locus. We designed the equation to eliminate the effects from the genome interaction. (a) We used the SCA-seq to establish the genome interaction table between genome bins. In genome interaction table, the open-open contacts were defined as the 2, which means the contacts could initiate the activation. The close-close contacts were defined as -2 and the close-open contacts were defined as 0. Then these values were summarized into the genome interaction table as means. The 3D genome interaction could amplify the original accessibility and form 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 could potentially activated the contacted sites, which we recognized as the “power center”. We plotted the expected chromatin accessibility from WT and CTCF-/-. The expected chromatin accessibilities were significantly higher in CTCF-/- than in WT (*t.test p<2.6e-16*), implying the CTCF-/- have more power centers to affect the contacted sites. We then plotted expected chromatin accessibility with the observed chromatin accessibility. The differences between the expected chromatin accessibilities and

the observed chromatin accessibilities (slope) were larger in CTCF-/- than in WT. This result suggested the chromatin accessibilities in CTCF-/- are more susceptible to the genome interaction. We selected an representative CTCF motif with the high expected signals. Each line represented the fragments in the concatemers. In CTCF-/-, the CTCF motif activated 50% contacted sites.

1. Lee, I. et al. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nature Methods* **17**, 1191-1199 (2020).
2. Durand, N.C. et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98 (2016).
3. Ulahannan, N. et al. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. *bioRxiv*, 833590 (2019).
4. Bell, A.C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**, 482-485 (2000).
5. Hark, A.T. et al. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**, 486-489 (2000).
6. 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**, 853-856 (2000).