CDCA-seq resolve the different chromatin structure on integral circular DNAs

Weitian Chen[1,2,*], Zhe Weng[2,*], Zhe Xie[2,3], Yeming Xie [2],Chen Zhang[2], Zhichao Chen [1,2], Fengying Ruan[2], Juan Wang[2], Yuxin Sun[2], Yitong Fang [2], Mei Guo [2], Yiqin Tong [2], Yaning Li [2], Chong Tang[2,4]

[1.] BGI Education Center，University of Chinese Academy of Sciences，Shenzhen 518083，China

[2.] BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China

[3.] Department of Biology, Cell Biology and Physiology, University of Copenhagen 13, 2100 Copenhagen，Denmark

[4.] Nantong University, Nantong, China, 226000

[5.] Nephrosis Precision Medicine Innovation Center, University of Beihua School of Medicine, Jilin 132011, China

[*] These authors contributed equally to this work.

Keywords: ecDNAs, chromatin accessibility, methylation, m^{6}A, methyltransferase

Running title: comprehensively chromatin accessibility on ecDNAs

Correspondence:

Chong Tang

Director of technology, BGI Shenzhen, China

Phone: 8618025420976

Email: tangchong@bgi.com

Abstract

Although ecDNAs have been a subject of sustained research activity for some years, the underlying mechanism driving the ecDNAs tumorigenesis has begun to unfold recently. Overall, from the results presented in conventional research, the high throughput short reads sequencing largely ignores the epigenetic status on most ecDNA regions except the junctional areas. We developed a method named CDCA-seq by using methylase to label the open chromatin without fragmentation, and exonuclease to enrich the ecDNA sequencing depth, followed by the long-read nanopore sequencing. Using this technology, the significantly different patterns of nucleosome/regulator binding were observed in ecDNAs at single-molecule resolution. These results further the understanding of the different regulatory mechanism on ecDNAs.

Introduction

Cancer is a malignant disease that is difficult to cure for various reasons, such as oncogene amplification, tumor evolution, and genetic heterogeneity [1-3]. These have been important topics of study in the literature for many years. Recently, it has been demonstrated that extrachromosomal DNA is closely related to carcinogenesis, such as promoting oncogene amplification[4], driving tumor evolution, and genetic heterogeneity[5-7]. Among extrachromosomal DNAs, extrachromosomal circular DNA (eccDNA), or microDNA, is the DNA that is arranged next to chromatin in a circular structure, usually very small in length (< 1 KB). In contrast, circular extrachromosomal DNA (ecDNA) may be cancer-specific DNA with an average size of 1.3 MB[8]. Although ecDNA was discovered very early, understanding it has been slow due to early techniques [9].

Many of these techniques are already well-known to study the ecDNA in genetic and epigenetic fields. Some attempts have been made to solve the ecDNA identification in sequencing data by improved algorithms. Most algorithms relied on the detection of ecDNA junction sequence, such as Circle-Map [10], AA (ampliconarchict) [11], CIRC_finder[12] , and others [13]. With the aid of these methods, the ecDNA was identified in numerous cancer tissues [12, 14, 15] [16] [5], aging cells [17], plasma [18, 19], and healthy somatic tissues [20]. Due to the rareness of ecDNAs in sequencing data, these approaches require an optimal solution to enrich the ecDNAs, which obtained circular DNA by digesting linear DNA

with nucleases, followed by rolling circle amplification [21]. To further improve the accuracy of finding ecDNAs, the long-read sequencing technology has also been used to verify the ecDNA junction structure [22] [23]. Besides the ecDNAs identification, future advancements are expected to result in the functional epigenetic study of ecDNAs. Given the rising prevalence of the ecDNAs and oncogene expression, an essential need is to know the chromatin status and transcription status of the ecDNAs. The most recent and advanced theory proposed by Wu et al. offers new insight for the highly accessible chromatin and high oncogene expression on ecDNAs by ATAC-seq, Chip-seq, and ATACsee [4]. Beyond that, limited studies are available about ecDNA epigenome study because it is still tricky to analyze its junction structure and epigenome information simultaneously.

Since eccDNAs possess a unique junction sequence pattern, its epigenetic information can be revealed by locating the neighboring junction regions without considering distal region, which was thought to be indistinguishable from linear genome sequences. However, the need for distal coverage of ecDNA epigenome arose due to the issues with current methods of short reads sequencing and short fragmentation in ATAC-seq [24] and MNase-seq [25]. Our research is also motivated by existing methods for assessing chromatin states using long-read sequencing, such as nanoNOMe-seq [26], SMAC-seq [27],fiber-seq[28]. We used EcoGII, the N6-methyladenosine (m^6A) methyltransferase, to soft label chromatin

accessible regions without fragmentation (Circular DNA Chromatin Accessibility Sequencing, CDCA-seq). We enriched the ecDNAs by using nuclease to digest the linear genome. The nanopore sequencing accurately detected the m^6A probed chromatin accessible regions and junctional structure properties simultaneously on ecDNAs in a long-range. Using this method, we found the high diversity of ecDNAs chromatin accessibility and coordination with distal regulators in single-molecule resolution, which has not been reported before.

Results

CDCA-seq comprehensively maps chromatin accessibility and nucleosome positioning in ecDNAs at multikilobase scale

The ecDNAs play a vital role in tumorigenesis due to their high chromatin accessibility and amplified oncogene expression [4]. Conventional approaches to study chromatin accessibility are based on the concept that the chromatins protected the bound sequence without being attacked by a transposon (Figure 1A) or MNase [25]. In the ATAC-seq, the transposon preferably tagmentated the openly accessible genome region, followed by the next-generation sequencing (NGS) (Figure 1A). However, these are not employed in most integral ecDNA chromatin studies due to the homologous ecDNA/genome sequences, making the problematic distinction between ecDNA and linear genome. In general, prior work in ecDNA

chromatin studies, based on NGS short reads, only observed the chromatin status in the junction region (200bp around junction) and took minimal consideration of the other distal areas on ecDNAs owing to limitation of techniques (>200bp to junction regions) (Figure 1A). We builted a generalized framework for solving problems in the integral ecDNA chromatin studies on the concept of the SMAC-seq[27] and fiber-seq[28]. We applied the soft labeling with EcoGII, the m^6A methyltransferase, preferably methylating the adenosine on the openly accessible DNA region without fragmentation by a transposon (Figure 1A). To improve the eccDNA capturing efficiency, the exonuclease was introduced to remove the linear genome DNA[29]. The integral ecDNAs could be sequenced by nanopore, and the probed m^6A could be detected [27]. In the following data analysis, we first identified the ecDNAs by head-to-tail junction locations with dynamically mapping the segments of sequences to the genome (Figure 1B). Based on the head-to-tail junction locations, we then reassembled the partial ecDNA sequences as the new reference and identify the m^6A signal based on the reassembled ecDNA sequence to prevent signal bias occurring in the junction region (Figure 1B).

Fundamental read length statistical analysis showed that the mean read length is around 10kb up to 100kb, which is 50x broader than the junctional region observed in conventional ATAC-seq[4] (Supplemental Figure 1). The long-read feature also makes the nanopore method optimal for applications such as SV, CNV, ecDNA identification with better sensitivity and specificity [30]. As expected, 80% ecDNAs

detected in our CDCA-seq could be validated through PCR (Supplemental Figure 2). ecDNAs and residual linear DNAs account for 0.9% and 99.1% in the total sequencing reads (Supplemental Figure 3) with exonuclease treatment. The m^6A possibility distribution in Megalodon, showed two distinct peaks for the treated sample. The distribution of the narrow peak with lower m^6A probability (mean=0.49) is similar to the background noise distribution (Supplemental Figure 4). Therefore, we set m^6A methylation probability over 0.53 as the cutoff for true m^6A signal. (Supplemental Figure 4). The real positive cutoff value was set as 0.53, and the m^6A calling specificity and sensitivity is 0.99 and 0.92 (Supplemental Figure 4). The residual linear DNA was used as an internal control to validate against the published ATAC-seq data[31]. The CDCA-seq achieved consistency and coherence with ATAC-seq data in various resolutions (Figure 1C, Supplemental Figure 5). Good concordance was also found when comparing our results against the published methods [27, 28]. The m^6A labeling deviation has been strongly reduced to 0.015 when accounting for the m^6A mean ratio (Supplemental Figure 6). The impact of exonuclease treatment and reproducibility has been also investigated here. (Supplemental Figure 7). These characteristics of CDCA-seq are critical for effectively measuring the chromatin accessibility on linear and circular DNAs in the multikilobase range.

Another remarkable feature of CDCA-seq is the single molecular resolution chromatin status on ecDNAs. At the single molecular level, the single base m^6A

possibility varied from 0.6 to 1 other than the genome level, whereas the genomic level, the m^6A possibility could be improved by multiple covered reads (Supplemental Figure 8). In practice, the resolution of chromatin accessibility is around 200bp. We adopted a Bayesian procedure to aggregate methylation probabilities and derived the accurate single-molecule accessibility calls over arbitrary size windows (Supplemental Figure 8). In summary, the CDCA-seq offers attractive features in terms of integral ecDNA chromatin status in the multikilobase range at the single molecular resolution.

The diversified chromatin accessibility on ecDNAs

Evidence from other studies by ATAC-seq and Chip-seq suggests that the active chromatin status and highly accessible chromatin on ecDNAs may be associated with the high transcription level of oncogenes [4]. To distinguish the ecDNAs molecules with linear DNA molecules in ATAC-seq and Chip-seq, we first need to screen out the short reads (~200bp) spanning the nonhomologous end-joining sequence of ecDNAs. One problem with these approaches is the potential bias to neglect the distal regions due to focusing on the reads of (~200bp) neighboring junctional sequence of ecDNAs. The long-read technology CDCA-seq may offer advantages for precisely detecting ecDNAs [30] [32] [22] and observing the distal chromatin status on integral ecDNAs. We observed an extensive catalog of 12997 different ecDNAs formed from chromosomal breakpoints between 0.05kb and up to

100kb (Supplemental Table 1). Gene ontology (GO) analysis of these ecDNA carried genes identified some significantly enriched GO terms, including GTPase-related activity, channel activity, nucleoside-triphosphatase activity, which play essential roles in cancer progression [33] (Supplemental Figure 9). By RNA-seq data analysis[34], the highly expressed ecDNA carried genes (25% rank), medium expressed genes (25~75% rank), and lowly expressed genes (75%~100% rank) are counted as 340, 464, and 589 respectively, indicating that not all the ecDNA carried genes are highly expressed.

By comparing the average chromatin accessibility between the ecDNAs and homologous linear DNAs, we found that the overall chromatin status of ecDNAs is 2x more accessible than that of linear DNAs (Figure 2A). These findings reinforce the general belief that the ecDNA amplification resulted in higher oncogene transcription [4], coupled with the enhanced chromatin accessibility in the junctional region. The data were subjected to the detailed mapping of the ecDNA chromatin status. We found that the chromatins in the ecDNA junctional areas are significantly more accessible than in other linear homologous regions (Figure 2B). This is an interesting find, as it suggests that the conclusion drawn by only observing the junctional areas of conventional ATAC-seq may be bias for the whole ecDNA chromatin. We calculate the average m^6A methylation fraction covering from the TSS (gene transcription start site) to the TES (gene transcription end site) on each gene-spanning read. A pairwise scatter plot of the average accessibility between

ecDNAs carried genes and linear genome had genes showed that 63% of the gene regions are more accessible on ecDNAs than linear DNAs (Figure 2C). Comparing the ecDNA and linear DNA chromatin profiles around the TSS/TES (+/- 500) revealed a significant difference in nucleosome depletion/occupancy patterns (Figure 2DE). The order nucleosome organization may impact access to ecDNA (Figure 2DE). Considering that the 63% gene regions are more accessible on the ecDNAs than on the linear DNA, we further plotted the chromatin structure around TSS/TES (+/-500bp) of these genes (Supplemental Figure 10). The formation of nucleosome depletion regions (NDRs) on linear DNAs is restricted to 200bp ahead TSSs. In contrast, the NDRs on ecDNAs are distributed in order (Supplemental Figure 10). The other 37% gene regions are more accessible on linear DNAs than ecDNAs. The TSSs/TESs (+/- 500bp) are also significantly more accessible on linear DNAs than ecDNAs with different NDR patterns (Supplemental Figure 11). The formation of large NDRs is restricted to TSSs on linear DNAs, which is not observed on ecDNAs.

Another illustration of the complex interplay between chromatin states on ecDNAs and linear DNAs relates to the transcriptional activity. The linear DNA chromatin state on the active genes (top 25% rank) largely devoid of nucleosomes on the TSSs due to extremely high transcription activity (Supplemental Figure 12). In contrast, the ecDNA chromatin structure of active genes adopts the distinct conformation, implying the different regulatory mechanisms on ecDNAs

(Supplemental Figure 12). For the transcriptionally inactive genes, the stationary nucleosome states are shown on the linear DNAs (Supplemental Figure 13). In contrast, the ecDNAs still have the active nucleosome organization on 300bp ahead TSSs, suggesting that chromatin accessibility is necessary but not sufficient for enhancer or promoter activity on ecDNAs (Supplemental Figure 13). In conclusion, the ecDNAs and linear DNAs have a significantly distinct nucleosome depletion/occupancy pattern in various conditions, suggesting the different gene regulatory mechanisms on between ecDNAs and linear DNAs.

The chromatin status on ecDNA and linear genome on the single-molecular resolution

The conventional ATAC-seq is based on statically calling the peak of the enriched read in a specific region[24]. Recent single-molecule and single-cell measurements of accessibility suggest that ATAC-seq on cell populations represent an ensemble average of distinct molecular states [35]. An essential attribute of the CDCA-seq measures the ecDNA chromatin accessibility in single molecular resolution by taking the advantages of the small variance (Supplemental Figure 6) and increased cumulative possibility in segments (Supplemental Figure 8). Measuring chromatin accessibility of the single linear DNA has also been done in the SMAC-seq [27] and fiber-seq [28].

We then asked whether the CDCA-seq could reveal multiple chromatin accessibility states on ecDNAs. The chromatin structure on linear DNAs

(chr10:42383201~42389251) adopts two distinct conformations; an inactive nucleosomal state and a state largely devoid of nucleosomes due to extremely high transcription activity [36] (Figure 3A).It was conventionally believed that the active nucleosome status occupied the majority in cancer cells. As expected, the 70% proportion of ecDNAs molecule comes from the very active chromatin state (Figure 3A). We observed the highly heterogenous nucleosome depletion/occupancy pattern on ecDNAs, and most chromatin molecules are not very active in the positive strand, suggesting the different transcription regulation on ecDNAs (Figure 3B upper panel). The different phenomena were also observed in other regions (Supplemental Figure 14). To avoid the bias conclusion drawn by the methylase heterogeneous activity, the other upstream and downstream region is chosen as quality control.

To further explore the limit of CDCA-seq's resolution limits, we studied methylation patterns in more detail. We next quantified strand-specific DNA accessibility and observed a strand-asymmetric DNA accessibility pattern in the linear genome (Supplemental Figure 14). The strand-asymmetric DNA accessibility pattern is also observed in ecDNAs, and both strands display high heterogeneity (Figure 3B, supplemental Figure 14). This strand-specific heterogeneity in methylation potential within the nucleosome may inform how transcription factors might interact with nucleosome-associated DNA in vivo.

Wu et al. show ecDNA enables ultra-long-range chromatin contact, permitting distant interactions with regulatory elements [4]. We next examined co-accessibility patterns in the ecDNAs and linear genome by assessing nucleosome positioning correlations. The nucleosomes have a higher correlation on ecDNAs than linear DNAs (Figure 4AB, Supplemental Figure 15). Moreover, the ecDNAs and linear DNAs adopt significantly different chromatin co-accessibility patterns (Figure 4AB, Supplemental Figure 15). Average co-accessibility profiles on linear DNAs reveal the detectable correlation between nucleosome positions up to two to three nucleosomes away. The ecDNAs correlation may be further and up to 20 nucleosomes away (Figure 4AB, Supplemental Figure 15). These results agree with the HiC result [4], that the ecDNAs have the distant chromatin interaction. It was interesting to note that the ecDNA demonstrates some ultra-distant anticorrelated states. Overall, the ecDNAs have high heterogeneity in each molecule and remote chromatin interaction, suggesting the different regulation mechanisms from linear DNAs.

Discussion

An understanding of ecDNAs may prove to be essential for tumorigenesis [5-7]. Most of the community's efforts have gone into solving ecDNA identification in various cancer tissues [12, 14, 15] [16] [5]. There has been an increasing research effort, and specialization in ecDNA open chromatin status to resolve the oncogene amplification on ecDNAs[4]. However, most studies focused on the short sequencing reads with junctional sequences detected to avoid the false-positive identification of ecDNAs and precisely determine the ecDNA epigenetic

status. A large subgroup (38%) and a large part (60%) of the ecDNAs covered regions of DNAs that are not unique in the reference genome, which is difficult to be discriminated [37]. In this study, we use nanopore sequencing to evaluate integral ecDNA chromatin accessibility on long strands of ecDNAs by applying m^6A methyltransferase to label open chromatin without fragmentation. Consistent with the findings of the other [4], 63% of the ecDNA carried genes have more accessible chromatin structure than the linear DNA. However, other 37% ecDNA chromatin on gene regions are less accessible than corresponding linear DNAs. Notably, the nucleosome depletion/occupancy patterns are significantly different between ecDNAs and linear DNAs. This single-molecule resolution method allows footprinting of protein and nucleosome binding and determination of epigenetic signature on chromatin accessibility. It is hoped that this study will lead to the new insight of comprehensively understanding ecDNA epigenome regulation.

Move to result section, our study would treat sample DNAs with an exonuclease that would remove most of the linear DNA molecules in the sample and increase the sequencing depth for the ecDNAs (0.9%). Few identified linear DNAs maybe come from ecDNA homologous regions without junctions, and the likelihood is around 0.9%, which is neglectable. Comparing with non-digest direct sequencing, we only get 0.1% ecDNA related reads (Supplemental Figure 16). The circular eccDNA enrichment fold is 10x. The exonuclease treatment not only improve the ecDNA sequencing coverage, but also the ecDNA detection specificity (Supplemental Figure 2). However, the DNA purification process would damage large-sized DNAs,

especially for the ecDNAs larger than 1MB [38]. The damaged ecDNAs could be digested during exonuclease digestion and missed in the sequencing. A method that gently purifies the large DNAs is preferable in further mega ecDNA studies.

A newer software Megalodon of m^6A signal calling is chosen and compared with Tombo, which is presented in other literature [27] [28]. In the ecDNAs m^6A calling, the Tombo would ignore half of the sequence or lose most ecDNAs for unknown reasons (Supplemental Figure 17). The Tombo lost 83% sensitivity on the ecDNA m^6A signal calling. Although the Megalodon improves the sensitivity on ecDNAs m^6A calling, it does not address the issue of the false-positive m^6A signal. Most adenosine could be recognized as m^6A with the possibility of 0.4~1 in Megalodon, other than the Tombo, which has higher specificity. The only known way to solve this problem is utilizing data training with negative control samples (Supplementary Figure 4). We used 0.53 as m^6A possibility cutoff, successfully discriminating the m^6A and false-positive with sensitivity 0.92 and specificity 0.99. In general, the Megalodon has demonstrated better performance in ecDNA research and as well as improved the specificity with our data training.

In the sequencing data, we found that the methylated treated DNA generated more data than the non-methylated DNAs, which is not consistent with SMAC-seq/fiber-seq[27] [28]. The highly open chromatin with high methylated sites may be enriched in our method. In our lab experiments, we found the heavily modified DNA is more resistant to exonuclease digestion, leading to the enrichment

of modified DNAs. The non-treated sample showed the lower overall methylation level (Supplemental Figure 18). However, the nucleosome occupancy positions were not significant affected by the exonuclease treatment (Supplemental Figure 19). Moreover, in the strand-specific view, the reverse strand reads are generally less abundant than the positive strands. This may also be due to the different methylation statuses on the positive and negative strands, which lead to different digestion effectiveness. This problem is usually overcome by increasing sequencing depth and normalization methods. We also suggest sequencing both treated and non-treated samples for ecDNA sequencing coverage and further improved quantification accuracy.

Only 63% of highly chromatin-accessible gene regions are observed in our experiment. However, Wu, etc., showed that the ecDNAs are overall highly chromatin accessible by ATACsee technology [4]. When comparing results from all other regions against the published data, good agreement was found that 80% of areas are highly accessible on the ecDNAs (Supplemental Figure 20). Most of the highly chromatin-accessible areas are distributed in the intron and intergenic regions (Supplemental Figure 21). The reasons for this remain unclear, but our results indicate that the ecDNAs have a highly open chromatin structure, especially in the intergenic and intronic regions.

The CDCA-seq provides the useful functionality to study the chromatin status on the integral ecDNAs, offering a deep insight into the different regulation mechanism of ecDNAs. The ecDNA enrichment step is to resort to using exonuclease treatment, causing the loss of mega ecDNAs. It is assumed that future advances will help to address these problems by DNA damages in the purification and sequencing depth.

Method

Cell culture

Human mammary gland carcinoma cell line MCF-7 was obtained from ATCC. MCF-7 were grown in DMEM(Gibco,11995065) supplemented with 10% FBS (Gibco,10099141), 0.01mg/ml insulin(), and 1% penicillin-streptomycin(Gibco, 15140122). The cell line was regularly checked for mycoplasma infection (Yeasen, 40612ES25).

Nuclei isolation and MTase treatment

Cells were grown to 70-80% confluency, and were collected by TrypLE (Gibco,12604013). After 300xg centrifuge for 5 minutes, nuclei were isolated with lysis buffer (100 mM Tris–HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1 mM EDTA, 0.5% CA630) for 5 minutes on ice. Nuclei were centrifuged at 300xg in wash buffer (100 mM Tris–HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1 mM EDTA) at 4 degree, and washed twice for 5 minutes and counted. 1x10^6 intact nuclei were subjected to an m^6A methylation reaction mixture containing 1x Cutsmart buffer (NEB), 200U of non-specific adenine methyltransferase M.EcoGII (NEB,

M0603S), 300mM sucrose, and 96 uM S-adenosylmethionine (NEB, B9003S) in 500ul volume. The reaction mixture was set up at a 37-degree thermomixer with shaking at 1000rpm for 30 minutes. S-adenosylmethionine was replenished at 640uM every 7.5 minutes at 7.5, 15, and 22.5 minutes into the reaction mixture. The reaction was stopped by adding an equal volume of stop buffer (20 mM Tris-HCl pH 7.4, 600 mM NaCl, 1% SDS, 10 mM EDTA). No methylation controls were treated in the same conditions without adding M.EcoGII in the reaction mixture. The samples were then treated with 20ul of Proteinase K (20mg/ml) at 55 degrees overnight, and the DNA was extracted with phenol: chloroform extraction and ethanol precipitation.

ecDNA isolation, purification, and sequencing

ecDNA was isolated by Circle-Seq[21] method, which digested linear DNA with modifications. Briefly, 10ug of M.EcoGII treated DNA was subjected to a reaction mixture containing 1x plasmid-safe reaction buffer, 20U plasmid-safe ATP-dependent DNase (Lucigen, E3101K), 1mM ATP, and nuclease-free water was supplemented to a final volume of 100ul. The reaction mixture was incubated at 37 degrees for 7 days. Every 24 hours, the reaction mixture was replenished by adding 20U plasmid-safe ATP-dependent DNase, 1mM ATP, and 0.4ul 10X plasmid-safe reaction buffer. Digested ecDNA was purified with 1.8X AMpure XP beads (Beckman Coulter).

Purified ecDNA was prepared for nanopore sequencing by ligation kit LSK-SQK108(ONT). The samples were 10kb by Covaris G tubes, end-repaired and dA-tailed using NEBnext Ultra II end-repair module (NEB), followed by clean-up using 1.8X AMpure XP beads.

Sequencing adaptors and motor proteins were ligated to end-repaired DNA fragments using blunt/TA ligase master mix (NEB), followed by clean-up using 0.4x AMpure XP beads. 1ug adaptor-ligated samples per flow cell were loaded onto PRO-002 flowcells and run on PromethION sequencers for up to 72h. Data were collected by MinKNOW v.1.14.

Base-calling and Linear DNA methylation calling

Reads from the ONT data were processed using Megalodon V2.2.9, which used Guppy base caller to base-calling, and Guppy model config res_dna_r941_min_modbases-all-context_v001.cfg released into the Rerio repository was used to identify DNA m^6A methylation. Megalodon_extras was used to get per read modified_bases from the Megalodon basecalls and mappings results. To further explore the accurate threshold of methylation probability, a control sample with almost no m^6A methylation was used as background noise, and the Gaussian mixture model was used to fit the methylation probability distribution generated by Megalodon.

ecDNA calling

ONT Reads meet the following conditions were defined as ecDNA molecules performed by the inner mappy/minimap2 aligner [39]. (1) One segment (>1kb) of an ONT read was mapped to the genome at one site, and another segment (>1kb) was mapped to the genome at another site. (2) Two segments were mapped to the same chromosome. (3) Two segments were mapped to the same strand of the genome. (4) Two segments in a pair showed outward orientation.

Nanopore ecDNA methylation calling

Due to ecDNA special structure, the m^6A calling cannot be successfully performed by aligning to the reference genome, especially for junctional regions. The custom python script was used to assemble ecDNA reference genome sequences according to the table generated from the previous step. Considering that the read length might be longer than the ecDNA reference, the ecDNA reference was subsequently preprocessed by adding 10M N to the ends to increase the mapping efficiency. The downstream step is performed in a similar way as linear DNA methylation calling.

Annotation and methylation configuration

TES, TTS, CDS, and other gene elements were downloaded from UCSC Table Browser, And the gene elements were processed into 50bp bin for downstream analysis. Linear DNA and ecDNA were also processed to the size of 50bp bin and sliding for 5bp. The accessibility score over multi base-pair windows was calculated as methylation ratio = m^6A bases in all covered reads under bin/ adenosine bases in all covered reads under the bin.

RNA-seq data analysis

The RNA-seq data of MCF-7 was downloaded from the Gene Expression Omnibus (GEO) repository database with the accession number GSE71862. The gene expression was divided into three categories: high, medium, and low, representing 25%, 25%-75%, and 75% gene expression rank, respectively.

Co-accessibility assessment

To evaluate co-accessibility patterns along the genome, we applied COA as follows. Each chromosome in the genome was split into windows of size w. For each such window (i, i + w), we identified another window (j,j+w) such that the span (i,j,w) was covered by ≥N reads. For each single spanning molecule k, accessibility scores (A) in each bin were then aggregated and binarized as described above. The local co-accessibility matrix between two windows was calculated:

$$\mathrm{COA}_{\mathrm{i,j,w}} = \mathrm{mean}_{\mathrm{N}} \left(1 - \frac{|\mathrm{Ai,w} - \mathrm{Aj,w}|}{\mathrm{Ai,w} + \mathrm{Aj,w}}\right)$$

Data availability

The data that support the findings of this study have been deposited into CNGB Sequence Archive (CNSA) [40] of China National GeneBank DataBase (CNGBdb)[41] with accession number CNP0001299.

Acknowledgment

Funding: 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

C.T. designed and supervised the experiments. Z.W. and X.Z perform the lab experiments; W.T.C. performs the bioinformatics data analysis. All others joined the data analysis.

Competing interest

The authors declare no competing interests.

References

1. Gillies RJ, Verduzco D, Gatenby RA: Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. Nat Rev Cancer 2012, 12:487-493.

2. Ray Chaudhuri A, Callen E, Ding X, Gogola E, Duarte AA, Lee JE, Wong N, Lafarga V, Calvo JA, Panzarino NJ, et al: Replication fork stability confers chemoresistance in BRCA-deficient cells. Nature 2016, 535:382-387.

3. Turajlic S, Swanton C: Implications of cancer evolution for drug development. Nat Rev Drug Discov 2017, 16:441-442.

4. Wu S, Turner KM, Nguyen N, Raviram R, Erb M, Santini J, Luebeck J, Rajkumar U, Diao Y, Li B, et al: Circular ecDNA promotes accessible chromatin and high oncogene expression. Nature 2019, 575:699-703.

5. Turner KM, Deshpande V, Beyter D, Koga T, Rusert J, Lee C, Li B, Arden K, Ren B, Nathanson DA, et al: Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature 2017, 543:122-125.

6. Paulsen T, Kumar P, Koseoglu MM, Dutta A: Discoveries of Extrachromosomal Circles of DNA in Normal and Tumor Cells. Trends Genet 2018, 34:270-278.

7. Verhaak RGW, Bafna V, Mischel PS: Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. Nature Reviews Cancer 2019, 19:283-288 %@ 1474-1768.

8. Chiu RWK, Dutta A, Henssen AG, Lo YMD, Mischel P, Regenberg B: What is extrachromosomal circular DNA and what does it do? Clinical Chemistry 2020, 66:754-759 %@ 0009-9147.

9. Bailey C, Shoura MJ, Mischel PS, Swanton C: Extrachromosomal DNA-relieving heredity constraints, accelerating tumour evolution. Ann Oncol 2020, 31:884-893.

10. Prada-Luengo I, Krogh A, Maretty L, Regenberg B: Sensitive detection of circular DNAs at single-nucleotide resolution using guided realignment of partially aligned reads. BMC Bioinformatics 2019, 20:663.

11. Deshpande V, Luebeck J, Nguyen N-PD, Bakhtiari M, Turner KM, Schwab R, Carter H, Mischel PS, Bafna V: Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. Nature communications 2019, 10:1-14 %@ 2041-1723.

12. Kumar P, Kiran S, Saha S, Su Z, Paulsen T, Chatrath A, Shibata Y, Shibata E, Dutta A: ATAC-seq identifies thousands of extrachromosomal circular DNA in cancer and cell lines. Science Advances 2020, 6:eaba2489 %@ 2375-2548.

13. Chen FZ, You LJ, Yang F, Wang LN, Guo XQ, Gao F, Hua C, Tan C, Fang L, Shan RQ: CNGBdb: China National GeneBank DataBase. Yi Chuan= Hereditas 2020, 42:799-809 %@ 0253-9772.

14. Koche RP, Rodriguez-Fos E, Helmsauer K, Burkert M, MacArthur IC, Maag J, Chamorro R, Munoz-Perez N, Puiggròs M, Garcia HD: Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. Nature Genetics 2020, 52:29-34 %@ 1546-1718.

15. Paulsen T, Shibata Y, Kumar P, Dillon L, Dutta A: Extrachromosomal circular DNA, microDNA, without

canonical promoters produce short regulatory RNAs that suppress gene expression. bioRxiv 2019:535831.

16. Verhaak RGW, Bafna V, Mischel PS: Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. Nat Rev Cancer 2019, 19:283-288.

17. Hull RM, King M, Pizza G, Krueger F, Vergara X, Houseley J: Transcription-induced formation of extrachromosomal DNA during yeast ageing. PLoS Biol 2019, 17:e3000471.

18. Zhu J, Zhang F, Du M, Zhang P, Fu S, Wang L: Molecular characterization of cell-free eccDNAs in human plasma. Sci Rep 2017, 7:10968.

19. Kumar P, Dillon LW, Shibata Y, Jazaeri AA, Jones DR, Dutta A: Normal and Cancerous Tissues Release Extrachromosomal Circular DNA (eccDNA) into the Circulation. Mol Cancer Res 2017, 15:1197-1205.

20. Møller HD, Mohiyuddin M, Prada-Luengo I, Sailani MR, Halling JF, Plomgaard P, Maretty L, Hansen AJ, Snyder MP, Pilegaard H, et al: Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. Nat Commun 2018, 9:1069.

21. Møller HD: Circle-Seq: Isolation and Sequencing of Chromosome-Derived Circular DNA Elements in Cells. Methods Mol Biol 2020, 2119:165-181.

22. deCarvalho AC, Kim H, Poisson LM, Winn ME, Mueller C, Cherba D, Koeman J, Seth S, Protopopov A, Felicella M, et al: Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. Nat Genet 2018, 50:708-717.

23. Mehta D, Cornet L, Hirsch-Hoffmann M, Zaidi SS-e-A, Vanderschuren H: Full-length sequencing of circular DNA viruses and extrachromosomal circular DNA using CIDER-Seq. Nature Protocols 2020,

15:1673-1689.

24. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ: ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol 2015, 109:21.29.21-21.29.29.

25. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K: Dynamic regulation of nucleosome positioning in the human genome. Cell 2008, 132:887-898.

26. Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, Sadowski N, Sedlazeck FJ, Hansen KD, Simpson JT, Timp W: Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nature Methods 2020, 17:1191-1199.

27. Shipony Z, Marinov GK, Swaffer MP, Sinnott-Armstrong NA, Skotheim JM, Kundaje A, Greenleaf WJ: Long-range single-molecule mapping of chromatin accessibility in eukaryotes. Nat Methods 2020, 17:319-327.

28. Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA: Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science 2020, 368:1449-1454 %@ 0036-8075.

29. Gaubatz JW, Flores SC: Purification of eucaryotic extrachromosomal circular DNAs using exonuclease III. Anal Biochem 1990, 184:305-310.

30. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al: Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res 2017, 27:677-685.

31. He HH, Meyer CA, Chen MW, Jordan VC, Brown M, Liu XS: Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. Genome Res 2012, 22:1015-1025.

32. Møller HD, Mohiyuddin M, Prada-Luengo I, Sailani MR, Halling JF, Plomgaard P, Maretty L, Hansen AJ, Snyder MP, Pilegaard H, et al: Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. Nature Communications 2018, 9:1069.

33. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000, 25:25-29.

34. Barutcu AR, Lajoie BR, McCord RP, Tye CE, Hong D, Messier TL, Browne G, van Wijnen AJ, Lian JB, Stein JL, et al: Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. Genome Biol 2015, 16:214.

35. Klemm SL, Shipony Z, Greenleaf WJ: Chromatin accessibility and the regulatory epigenome. Nat Rev Genet 2019, 20:207-220.

36. Conconi A, Widmer RM, Koller T, Sogo J: Two different chromatin structures coexist in ribosomal RNA genes throughout the cell cycle. Cell 1989, 57:753-761.

37. Moller HD, Parsons L, Jorgensen TS, Botstein D, Regenberg B: Extrachromosomal circular DNA is common in yeast. Proc Natl Acad Sci U S A 2015, 112:E3114-3122.

38. Smith CL, Cantor CR: 6 - Purification, Specific Fragmentation, and Separation of Large DNA Molecules. In Recombinant DNA Methodology. Edited by Wu R, Grossman L, Moldave K. San Diego: Academic Press; 1989: 139-157

39. Li H: Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018, 34:3094-3100.

40. Guo X, Chen F, Gao F, Li L, Liu K, You L, Hua C, Yang F, Liu W, Peng C: CNSA: a data repository for archiving omics data. Database 2020, 2020.

41. Chen FZ, You LJ, Yang F, et al. CNGBdb: China National GeneBank DataBase, Hereditas. 2020;42(08):799-809. doi:10.16288/j.yczz.20-080

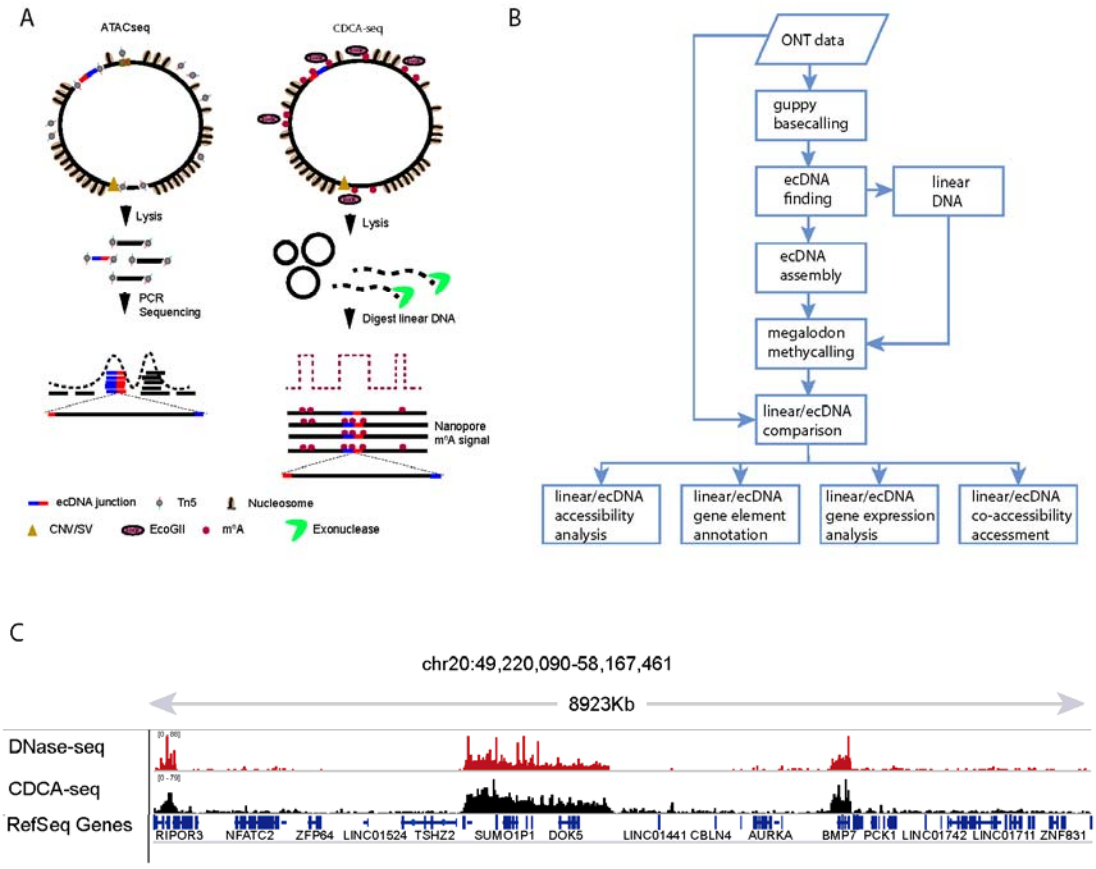

Figure1. The CDCA-seq for profiling chromatin accessibility and nucleosome position at ecDNAs. (A) Intact chromatin is treated with m^6A methyltransferase (EcoGII), which preferentially methylate DNA bases in open chromatin region on ecDNAs and linear DNAs. High molecular weight DNA is then isolated and subjected to exonuclease digestion to remove partially linear DNAs. The rest DNAs are subjected to nanopore library construction and nanopore sequencing. The data were aligned to genome to identify ecDNAs based on head-to-tail pattern. The methylated bases are used to reconstruct nucleosomes in ecDNAs and rest linear DNAs. In contrast, the ATAC-seq used the transposon to attack the open chromatin. The tagmentated short fragments are amplified and subjected to NGS. The short

reads are aligned with genome to identify ecDNAs based. The mapped reads are calling as peaks to represent the open chromatin region. (B) The bioinformatics pipeline of CDCA-seq. The signal data were processed through guppy basecalling to generate sequence. The sequences were aligned to genome to identify the linear DNA and ecDNAs. We assembled the ecDNA sequence reference. Based on the ecDNA and linear DNA reference, we used Megalodon to call the m^6A sites base on ecDNA and linear DNAs. Then we perform the accessibility analysis, gene element annotation, gene expression analysis, co-accessibility assessment. (C) Large aggregate CDCA-seq signal enrichments match closely with DNase-seq accessibility peaks. (Chr20:49,220,090-58,167,461)

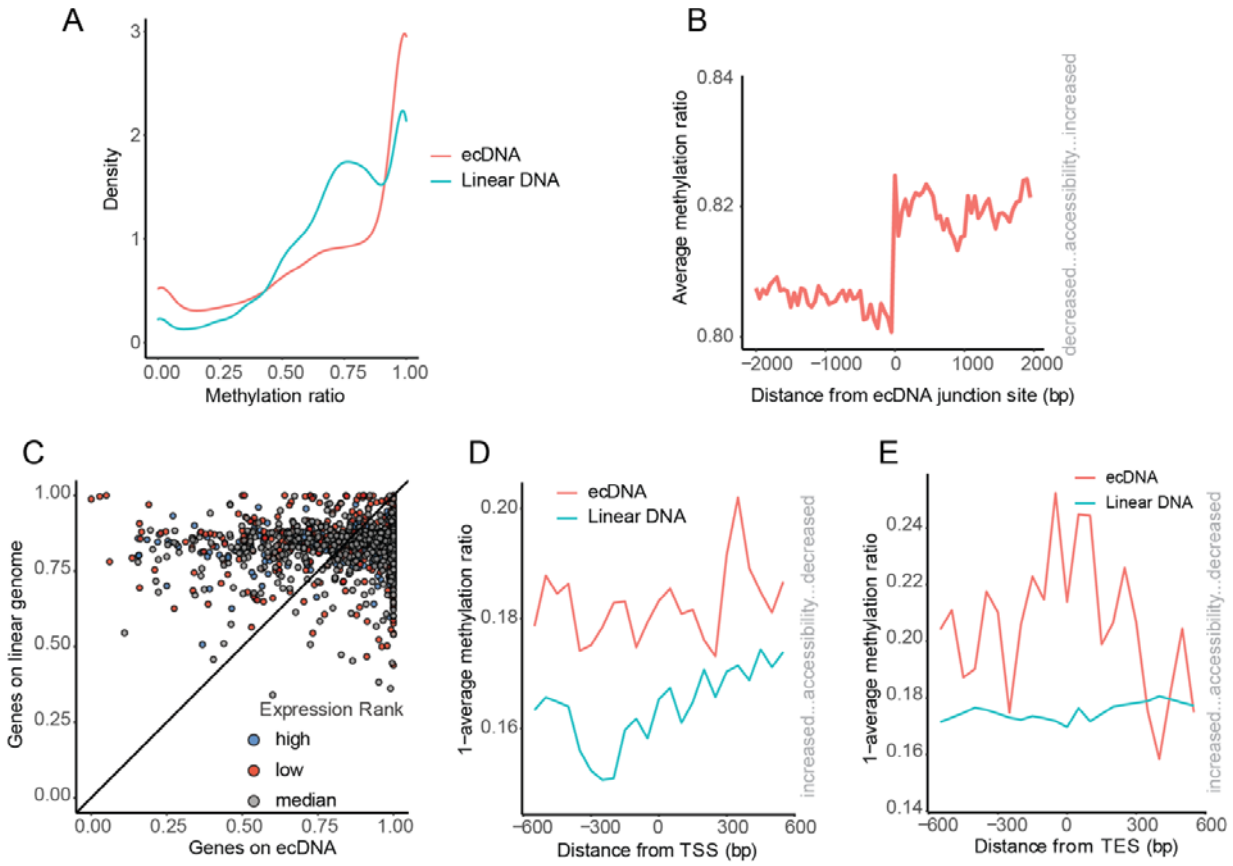

Figure 2. The ecDNA and linear DNA have the different chromatin accessibility pattern.

(A) The density distribution of the methylation ratio on ecDNAs and linear DNAs. (B) The average chromatin status around ecDNA neighboring regions. The junction site and its right neighboring regions demonstrate the more open chromatin. (C) The average methylations of gene regions on ecDNA and linear DNAs (from TSS to TES). The genes were classified as two groups: I. The genes on linear DNA have more open chromatin structure than ecDNA carried genes; II. The ecDNA carried genes have more open chromatin structure than the genes on linear DNA. (D) Average CDCA-seq profile around transcription start site on ecDNAs and linear DNAs. (E) Average CDCA-seq profile around transcription end site on

ecDNAs and linear DNAs. (aggregated over 50-bp windows sliding every 5 bp; the sequencing depth is normalized for ecDNA and linear DNA;see Methods for details)

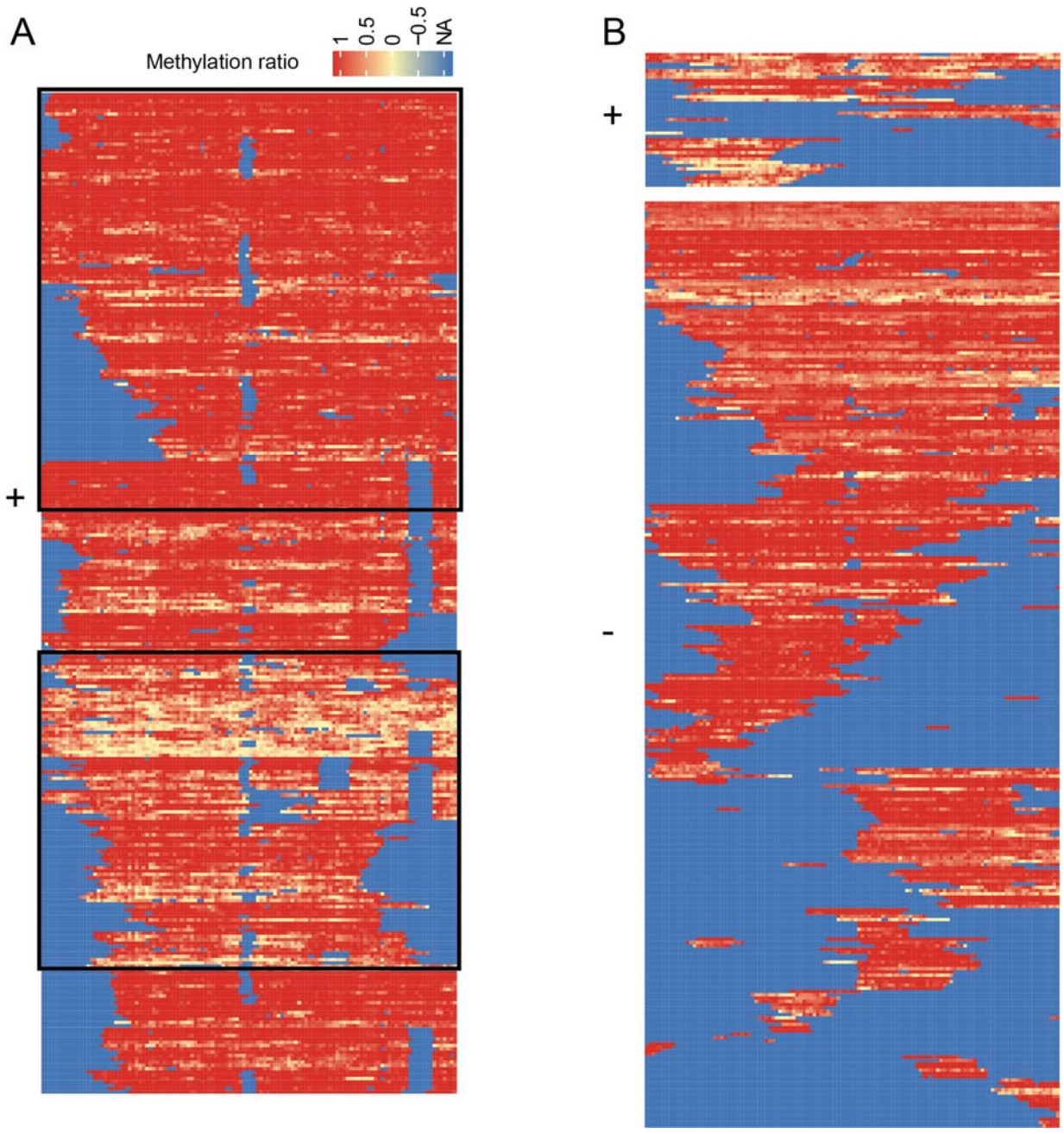

Figure 3. CDCA-seq reveals the distribution of alternative chromatin states of ecDNA arrays. A. Shown are all reads covering the linear DNA region chr10:42383201~42389201.The box highlights the active and inactive chromatins. B. Shown are all reads covering the ecDNA region chr10:42383201~42389201. The upper panel indicates the positive strand, and the lower panel indicates the negative strand.

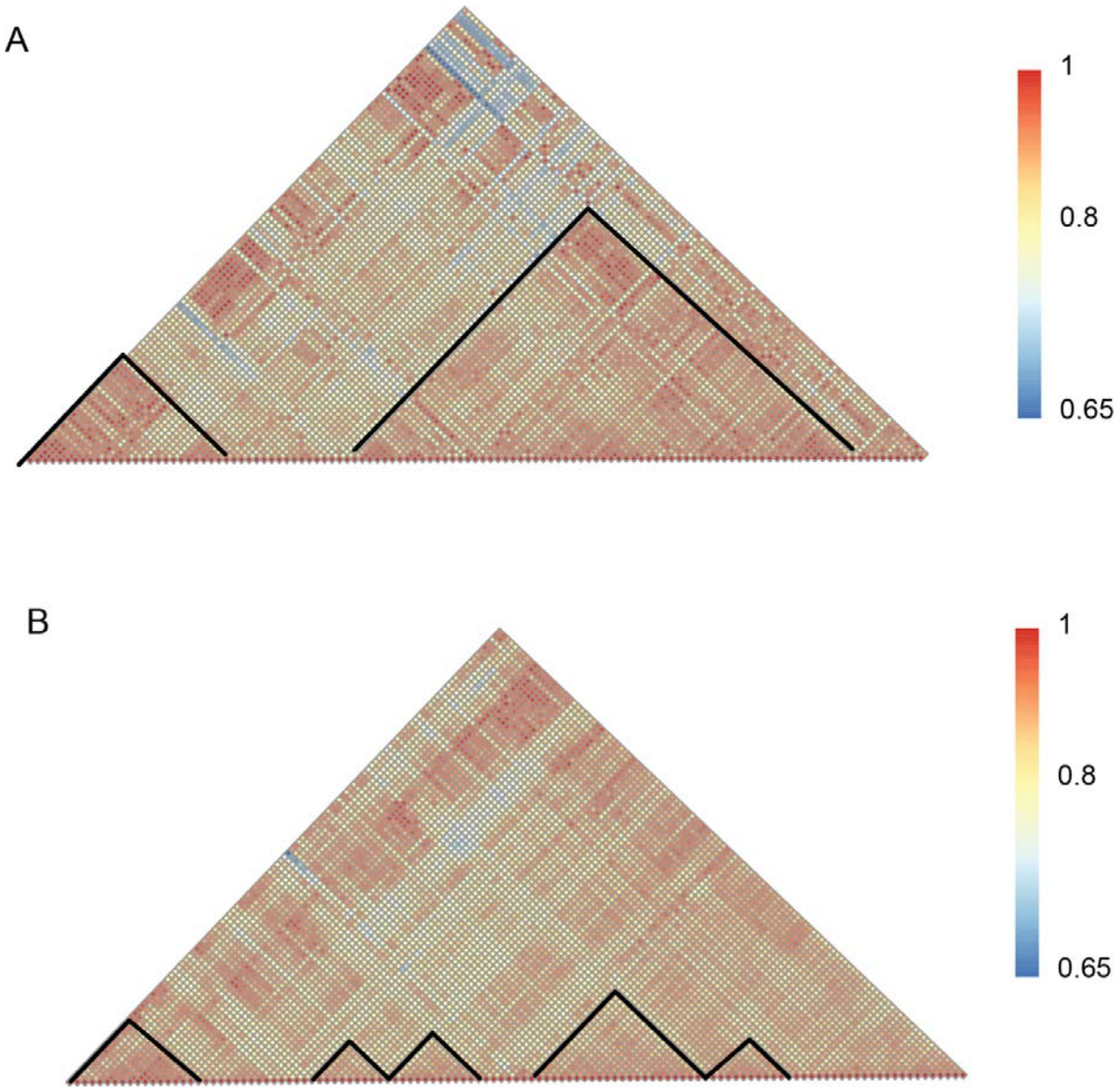

Figure 4. Chromatin co-accessibility profiles for the chr10:42383201~42389201 show correlation and anticorrelation in ecDNA and linear DNA. (A) Chromatin co-accessibility profiles for the chr10:42383201~42389201 show correlation and anticorrelation on ecDNA. Red indicates the positive correlation and blue indicates the anticorrelation. (B) Chromatin co-accessibility profiles for the chr10:42383201~42389201 show correlation and anticorrelation on linear DNA.