

METHODOLOGY

Open Access



Revealing long-range heterogeneous organization of nucleoproteins with 6mA footprinting by ipdTrimming

Wentao Yang^{1,2}, Xue Qing Wang^{1,2}, Fan Wei¹, Jingqi Yu^{1,3}, Yifan Liu^{1*} and Yali Dou^{1,2*}

*Correspondence:
Yifan.Liu@med.usc.edu;
Yalidou@usc.edu

¹ Department of Cancer Biology,
Keck School of Medicine,
University of Southern California,
Los Angeles, CA 90033, USA

² Department of Medicine, Keck
School of Medicine, University
of Southern California, Los
Angeles, CA 90033, USA

³ Present Address: University
of Chinese Academy of Sciences,
Beijing 100864, China

Abstract

Enabled by long-read sequencing technologies, particularly Single Molecule, Real-Time sequencing, N⁶-methyladenine (6mA) footprinting is a transformative methodology for revealing the heterogeneous and dynamic distribution of nucleosomes and other DNA-binding proteins. Here, we present ipdTrimming, a novel 6mA-calling pipeline that outperforms existing tools in both computational efficiency and accuracy. Utilizing this optimized experimental and computational framework, we are able to map nucleosome positioning and transcription factor occupancy in nuclear DNA and establish high-resolution, long-range binding events in mitochondrial DNA. Our study highlights the potential of 6mA footprinting to capture coordinated nucleoprotein binding and to unravel epigenetic heterogeneity.

Background

One of the central tasks in epigenetics is to map the distribution of nucleosomes and other DNA-binding proteins across the genome [1, 2]. This is usually achieved by isolating and sequencing DNA fragments corresponding to a specific nucleoprotein [3–6]. Various nucleases, such as DNase I, micrococcal nuclease (MNase), and Tn5 transposase, are most commonly used for DNA fragmentation in techniques like DNase-seq, MNase-seq, ATAC-seq [7–9], or CUT&RUN and CUT&Tag [10, 11]. The positional information of a nucleoprotein can be inferred from altered nuclease access, but the connectivity information between coordinate binding events on the same DNA molecule is lost. The derived data need to be aggregated to reveal statistically significant patterns, which obscure heterogeneous or dynamic protein-DNA interactions. The position for a nucleoprotein can also be registered by altered access to a DNA methyltransferase (MTase) and corresponding changes in DNA methylation (N⁶-methyl adenine: 6mA; 5-methyl cytosine: 5mC), implemented in techniques like DamID and NOME-seq [12, 13]. DNA methylation footprinting, when coupled with long-read sequencing technologies, such as Single Molecule, Real-Time (SMRT/PacBio) sequencing and Oxford



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Nanopore (ONT) sequencing, has the potential to reveal joint binding events and long-range organization on individual chromatin fibers [13–18]. Particularly promising is the combination of 6mA footprinting (6mA-FP) and SMRT sequencing. SMRT sequencing can call 6mA (as well as 5mC) on individual DNA molecules multi-kb in length [19, 20]. Since Mammalian genomic DNA lacks 6mA [21–24], so essentially all 6mA sites can be attributed to 6mA-FP. Combined with the ubiquitous presence of adenine sites and application of promiscuous 6mA MTases, 6mA-FP may detect DNA binding events at the single molecule level and with near-basepair resolution. Variably implemented as chromatin fiber sequencing (fiber-seq) [14], single-molecule adenine methylated oligonucleosome sequencing assay (SAMOSA) [16], and SAMOSA by tagmentation (SAMOSA-Tag) [25], 6mA-FP has been employed to study global chromatin organization, mitochondrial DNA compaction, and the epigenetic landscape. However, effective and reliable detection of base modifications, including 6mA, by long-read sequencing technologies remains a challenge and an area of active research [26, 27].

In SMRT sequencing, 6mA is distinguished from the unmodified adenine by its perturbation to DNA polymerase kinetics—specifically increase in the time between nucleotide incorporation, referred to as the inter-pulse duration (IPD) (Fig. 1A) [20, 28]. 6mA mapping is usually achieved at the ensemble/genomic level, by combining different DNA molecules covering the same genomic position to overcome random fluctuations in IPD [29–33]. To generate accurate base calling and IPD information at the single molecule level, multiple strand-specific readouts of the same DNA template, referred to as subreads, are combined in Circular Consensus Sequencing (CCS; also known as PacBio HiFi Sequencing) (Fig. 1A) [19, 20, 24, 28]. Even though several bioinformatic pipelines have been developed for 6mA calling with SMRT CCS data [14, 16, 34–36], various issues limit their widespread adoption, and consequently, effective implementation of 6mA-FP. We have previously developed a pipeline that processes the subread data from the PacBio Sequel systems and generates highly accurate 6mA calls at the single molecule level, but at a steep computational cost [34]. Furthermore, it is not applicable for the latest PacBio platforms (Revio and Vega), which no longer output the subread kinetics data [37]. Current 6mA calling methods, including Fibertools (applied in fiber-seq) [35], as well as Jasmine (just released by PacBio) [36], still need significant improvement for 6mA calling for broadly applicable, accurate, and cost-effective 6mA calling at the single molecule level.

Here, we developed a new ipdTrimming pipeline that matched or beat ipdSummary [34, 38–40]—the current gold standard in the field—for 6mA calling in all parameters, including background noise, recall, and precision, at only a fraction of its computation cost. Furthermore, this pipeline can be seamlessly integrated with the latest PacBio systems (e.g., Revio). Critically, across a diverse range of samples sequenced on both Sequel and Revio systems, ipdTrimming substantially outperformed Fibertools and Jasmine. Leveraging this cutting-edge bioinformatic workflow in conjunction with optimized 6mA-FP experimental procedures, we conducted detailed investigation of nucleoprotein organizations within human cells, encompassing both nuclear chromatin and mitochondrial nucleoid structures. Our analyses, at near-base pair resolution, revealed complex, long-range, and dynamic interactions involving nucleosomes, transcription factors/cofactors, and mitochondrial proteins. Our work positions 6mA-FP as a transformative

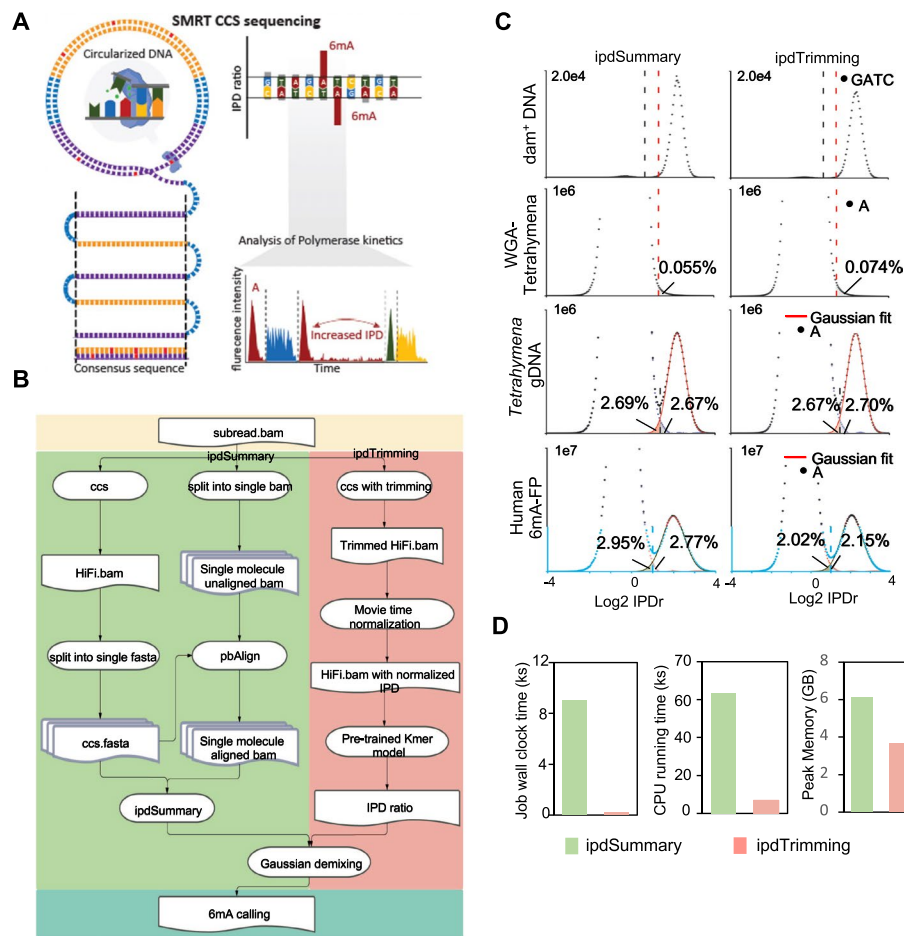


Fig. 1 A new 6mA calling pipeline for SMRT CCS. **A** Schematic for SMRT CCS and 6mA calling. **B** Flowchart for ipdSummary and ipdTrimming pipelines, both of which use the Sequel subread data for 6mA calling. **C** Performance of ipdSummary and ipdTrimming on four different datasets. IPDr (log transformed) distributions are plotted. The IPDr thresholds for calling 6mA are indicated by black lines. For dam⁺ plasmid DNA dataset, the thresholds for 98.5% 6mA recall are indicated by red lines. The same recall thresholds are used to calculate the background noise level for 6mA calling in the WGA dataset. False positive rates (FPR: right shaded area) and false negative rates (FNR: left shaded area) for both *Tetrahymena* gDNA and human 6mA-FP samples were calculated after Gaussian demixing of two IPDr distribution peaks. **D** Computation cost of ipdSummary and ipdTrimming, including job wall clock time, CPU running time, and peak memory usage

tool for mapping the intricate organization of nucleoproteins and uncovering diverse epigenetic states at the single-molecule level.

Results

Development of ipdTrimming as a new 6mA calling pipeline for SMRT CCS

IPD conversion is a critical step in kinetics analysis of the PacBio Sequel data: for each site on both strands of a DNA molecule, multiple subread IPD values are converted to a single CCS IPD value [20, 41, 42]. By default, subread IPD values are averaged to generate the CCS IPD value by the PacBio CCS module [37]. However, DNA polymerase pausing stochastically introduces very large subread IPD values [28, 39, 43]. These outliers can substantially shift CCS IPD values generated by the CCS module, leading to false

6mA calls. The PacBio ipdSummary module, used for kinetics analysis at the ensemble level [37], controls this background noise by globally capping very large subread IPD values [39]. We have previously adapted ipdSummary for single molecule level 6mA calling, with high accuracy but at a steep computation cost [34]. Here, we developed a greatly simplified computation pipeline by introducing local trimming into IPD conversion: for each site, the maximum subread IPD value is removed before the rest are averaged to generate the CCS IPD value using the modified CCS module (Fig. 1B). The CCS IPD value, after standard transformation [44], was then compared to a reference IPD value of its unmodified counterpart embedded in the same local sequence (referred to as the kmer model) [45]. The IPD ratio (IPDr) thus derived was used for 6mA calling (Fig. 1B).

We systematically compared the performance of our new pipeline, henceforth referred to as ipdTrimming, with ipdSummary on several benchmarking datasets: (1) a plasmid fragment from *dam*⁺ *E. coli*, with all A sites in GATC presumably methylated, serving as a positive control, (2) *Tetrahymena* whole genome amplification (WGA), with no base modifications, serving as a negative control, (3) native *Tetrahymena* genomic DNA, with ~1% of all adenine sites methylated, (4) 6mA-FP of in vitro assembled nucleosome arrays and human chromatin, with ~10% of all adenine sites methylated (Fig. 1C, Additional file 1: Fig. S1-S3, and Additional file 2: Table S1). A high cutoff for strand-specific subread counts increases 6mA calling quality, but at the expense of reducing sequencing coverage (Additional file 2: Table S2). As a compromise between the two opposing demands, we set a threshold of ≥ 10 subreads for selecting DNA molecules as input for 6mA calling, which ensures <0.05 false negative and false positive rates (Additional file 2: Table S2). For the first two datasets, IPDr values for all A sites were distributed predominantly within a single peak of either high IPDr corresponding to 6mA or low IPDr corresponding to unmodified A (Fig. 1C, Additional file 1: Fig. S1). The 6mA background noise level for the WGA dataset was very low using both ipdTrimming (0.074%) and ipdSummary (0.055%) (Fig. 1C, Additional file 1: Fig. S2). For native *Tetrahymena* genomic DNA and 6mA-FP datasets that had both 6mA and unmodified A sites, IPDr exhibited a bimodal distribution: a large peak of low IPDr corresponding to unmodified A and a small peak of high IPDr corresponding to 6mA (Fig. 1C). We resolved the 6mA and unmodified A peaks with Gaussian demixing and estimated the false positive and false negative rates (FPR and FNR) of 6mA calling (Fig. 1C, Additional file 2: Table S3). The performance of ipdTrimming was nearly identical to ipdSummary for the *Tetrahymena* genomic DNA dataset that has less than 1% overall 6mA level (Fig. 1C). The performance of ipdTrimming was noticeably better for the human 6mA-FP dataset that has >10% overall 6mA level, with FPR improving from 2.77 to 2.15% and FNR improving from 2.95 to 2.02% (Fig. 1C, Additional file 2: Table S3). The ipdTrimming also greatly outperformed the CCS-kmer pipeline using the default CCS module to generate CCS IPD values for 6mA calling (Additional file 1: Fig. S2, Additional file 2: Table S3).

Comparing to the endogenous 6mA sites previously identified by ipdSummary in *Tetrahymena* [34], we found that ipdTrimming was able to recall 99% of all 6mA sites (Additional file 1: Fig. S2). Furthermore, ipdTrimming was able to demonstrate all key features consistent with semiconservative transmission of 6mA in *Tetrahymena* [34], including (1) 6mA exclusively occurs at the ApT dinucleotide, (2) full-6mApT is predominant over

hemi-6mA_{pT}, and (3) hemi-6mA_{pT} is predominantly found in only one strand of newly synthesized DNA duplex (Additional file 1: Fig. S2).

We next examined 6mA-FP on a previously published 6mA-FP dataset for an array of in vitro assembled nucleosomes [16] as well as human chromatin (Fig. 1C, Additional file 1: Fig. S3, Additional file 2: Table S1, S4). In both cases, the vast majority of 6mA sites were jointly called by ipdTrimming and ipdSummary (Additional file 1: Fig. S3, Additional file 2: Table S3). The 6mA and unmodified A peaks were better resolved in ipdTrimming (Fig. 1C). Overall, 6mA calling by ipdTrimming was more conservative than ipdSummary (Additional file 1: Fig. S3). The 6mA is known to have secondary off-site effects on IPDr values, most prominently at positions 5 nucleotides upstream (5') of primary 6mA sites [46]. Intriguingly, ipdTrimming made substantially fewer 6mA calls at one of the likely secondary sites on the nucleosome array than ipdSummary while maintaining all key features, including the ~10 bp oscillation pattern (Additional file 1: Fig. S3).

We next evaluated the computation cost of ipdTrimming. The new ipdTrimming pipeline greatly streamlined data processing and transferring (Fig. 1B). As a result, it had much lower computation cost than ipdSummary, with job wall-clock time at 0.023 ×, CPU running time at 0.11 ×, and peak memory usage at 0.60 × (Fig. 1D). We conclude that the streamlined and open source ipdTrimming pipeline effectively supersedes the proprietary ipdSummary package for kinetic analysis of Sequel data, with room for future improvement.

Comparison of 6mA calling pipelines on Sequel and Revio platforms

We compared ipdTrimming with current 6mA calling methods for Sequel data, including Fibertools—the latest open-source pipeline tailored for 6 mA-FP (fiber-seq) [35], as well as Jasmine—a pipeline for calling base modifications (5mC and 6mA) recently released by PacBio [36]. With the default setting, both Fibertools and Jasmine had lower recall rates (86.1% and 77.7%, respectively) for 6mA in GATC sites in the positive control *dam*⁺ plasmid DNA (presumably 100% 6mA), compared to 98.7% using ipdTrimming (Fig. 2A, Additional file 2: Table S5). For the native *Tetrahymena* genomic DNA, Fibertools and Jasmine only recalled 73.7 and 78.5% of the 6mA sites that were called by ipdTrimming (Fig. 2A, Additional file 2: Table S5). They also falsely identified unmodified A sites as 6mA, with FPR at 11.2 and 10.1%, respectively (Additional file 2: Table S5). Close examination revealed that Fibertools and Jasmine made 6mA calls at all four ApN dinucleotides, instead of exclusively at the ApT dinucleotide as expected for *Tetrahymena* genomic DNA (Additional file 1: Fig. S4). For 6mA-FP of both in vitro assembled nucleosome arrays and human chromatin, Fibertools and Jasmine provided a general profile consistent with the beads-on-a-string nucleosome distribution (Additional file 1: Fig. S5). Nonetheless, there were substantial discrepancies between 6mA calls made by them and ipdTrimming due to much lower recall rates (Fig. 2A) and high FPR (Additional file 1: Fig. S5, Additional file 2: Table S5). Analysis using Fibertools and Jasmine led to diminishing of key features, such as the ~10-bp oscillation pattern of nucleosomal DNA (Additional file 1: Fig. S5). We conclude that for the Sequel data, the ipdTrimming pipeline greatly outperforms Fibertools and Jasmine in all metrics, establishing a new 6mA calling standard for both endogenous 6mA detection and 6mA-FP.

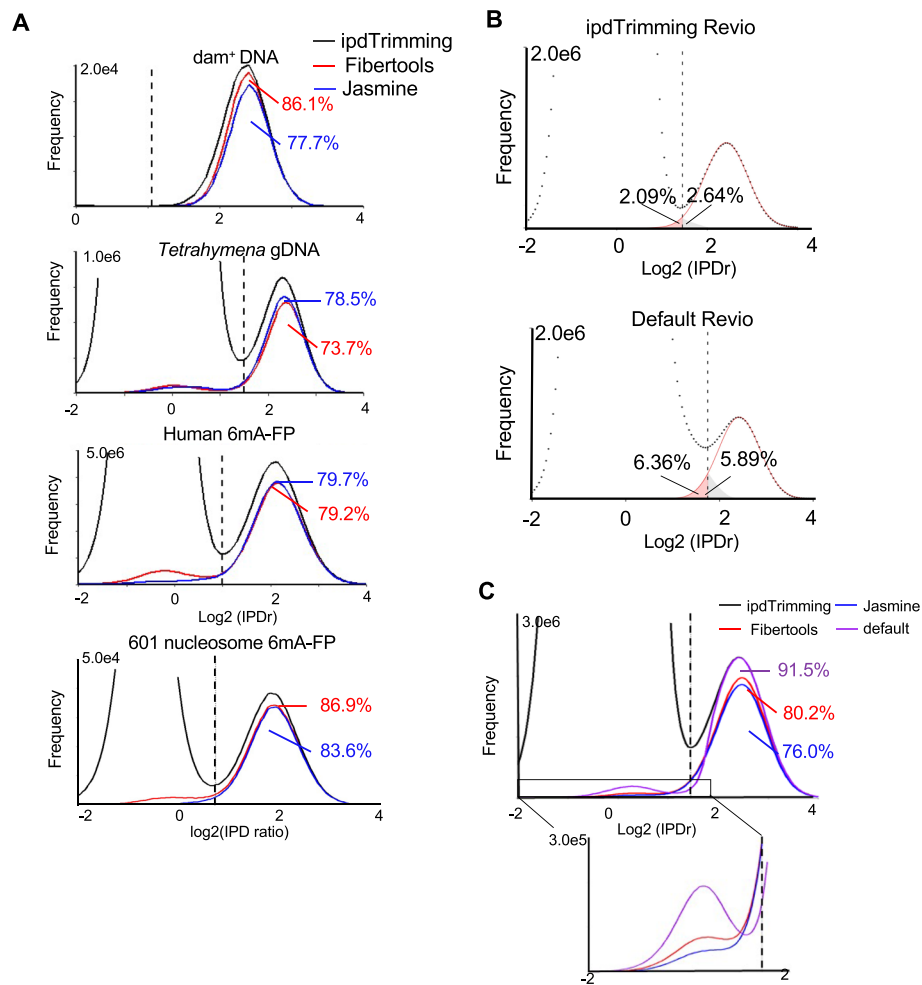


Fig. 2 Comparison with other 6mA calling methods. **A** Performance of ipdTrimming on the Sequel subread data, compared to Fibertools and Jasmine. For ipdTrimming, IPDr (log transformed) distributions are plotted for all A sites from samples of dam⁺ plasmid DNA, *Tetrahymena* gDNA, and 6mA-FP of human chromatin and Widom 601 nucleosomes. The IPDr thresholds for calling 6mA are indicated by black lines. 6mA sites called by Fibertools and Jasmine were mapped back to the same plot according to their ipdTrimming IPDr. **B** 6mA calling performance with the default and ipdTrimming for Revio data. IPDr (log transformed) distributions are plotted for all A sites in the ApT dinucleotide from a *Tetrahymena* gDNA sample sequenced on the Revio system. The IPDr thresholds for calling 6mA are indicated by black lines. False positive rates (FPR: right shaded area) and false negative rates (FNR: left shaded area) were calculated after Gaussian demixing of two IPDr distribution peaks. **C** Performance of ipdTrimming on the Revio subread data, compared to Fibertools, Jasmine, using the default pipeline. For ipdTrimming, IPDr (log transformed) distribution is plotted for all A sites in the ApT dinucleotide from a *Tetrahymena* gDNA sample. The IPDr threshold for calling 6mA is indicated by a black line. 6mA sites called by Fibertools, Jasmine, and Gaussian demixing using the default CCS IPD data (also the default Revio output) were mapped back to the same plot according to their ipdTrimming IPDr. Zoom-in: Distribution of probable false positive 6mA sites (with low ipdTrimming IPDr values) called by Fibertools, Jasmine, and the default pipeline

For the new high-throughput PacBio Revio platform, the on-instrument IPD conversion uses the default CCS module and the subreads are not routinely provided in output. To determine ipdTrimming performance on the Revio system, we obtained the Revio subread data for native *Tetrahymena* genomic DNA from the PacBio service laboratory. We evaluated ipdTrimming performance for 6mA calling using both the default and

modified CCS modules (Fig. 2B, Additional file 1: Fig. S6). When the IPDr distribution was plotted, the 6mA and unmodified A peaks were much better resolved by ipdTrimming using the modified CCS module as compared to the default module, improving FPR from 5.89 to 2.64% and FNR from 6.36 to 2.09% (Fig. 2B). The FPR and FNR for 6mA calling by ipdTrimming using Revio data largely matched their Sequel counterparts (Additional file 1: Fig. S2).

Using the 6mA_PT calls based on ipdTrimming of subread IPD values as the ground truth, we compared the performance of Fibertools and Jasmine on the *Tetrahymena* genomic DNA using the default Revio output. Fibertools and Jasmine had much lower recall rates at 80.2 and 76.0%, respectively (Fig. 2C, Additional file 2: Table S5). The FPR stands at 7.1% for Fibertools and 5.6% for Jasmine (Additional file 2: Table S5). Interestingly, ipdTrimming using the default Revio output maintained higher recall rate at 91.5%, although FPR was modestly elevated at 10% (Fig. 2C, Additional file 2: Table S5). We conclude that ipdTrimming using both the default and modified Revio output is superior to Fibertools and Jasmine on 6 mA recall rates. Inclusion of subread IPD values in Revio output can be very effective in controlling background noise introduced by polymerase pausing and, therefore, significantly reduces FPR in processing Revio data. Indeed, incorporating subread trimming in IPD conversion for Revio data from *Tetrahymena* genomic DNA greatly reduced FPR of not only ipdTrimming (Fig. 2B), but also Fibertools and Jasmine in 6mA calling (Additional file 2: Table S5).

Nucleosome position and dynamics revealed by 6mA-FP

We applied the new ipdTrimming pipeline to re-analyze the 6mA-FP dataset for an in vitro assembled nucleosome array (Fig. 3A, Additional file 1: Fig. S3, Additional file 2: Table S1, S4) [16], with an emphasis on the nucleosome features [14, 16], especially the fine patterns of 6mA distribution on nucleosomal DNA. Consistently with the previous study [16], linker DNA regions were densely methylated on individual DNA molecules, with many sites near saturation at the ensemble level (Fig. 3A, Additional file 1: Fig. S3). We aggregated multiple closely spaced 6mA sites into a 6mA cluster at the single molecule level; complementarily, we designated the interval between adjacent 6mA clusters a 6mA footprint if it contains multiple unmodified A sites (Fig. 3A, bottom). 6mA clusters readily expanded beyond the linker DNA boundary and into the 147 bp Widom 601 sequence constituting the nucleosome core particle (Fig. 3A–C) [47]. Most expansions stayed in the terminal 13 bp of the Widom 601 sequence and stopped outside the central 121 bp nucleosomal DNA (nucleosome dyad ± 60 bp), supporting an octasome footprint with unwrapping of DNA in contact with the H3 α N helix and the root of the H3 N-terminal tail [48, 49]. Some 6mA clusters encroached into the region ± 30 to ± 60 bp from the nucleosome dyad, supporting an octasome-to-hexasome transition with unwrapping of DNA in contact with the H2A-H2B dimer [49]. Intriguingly, we also found a single A site on the left half (-31 bp from the dyad) with usually high 6mA penetrance (Fig. 3B), supporting dynamic opening of the interface between the H2A-H2B dimer and the H3-H4 tetramer [50]. Very rarely did they reach within ± 30 bp from the nucleosome dyad, representing complete loss of H2A-H2B protection from either side or both. Notably, DNA unwrapping occurred preferentially from the right half of the Widom 601 nucleosome (Fig. 3B, C), consistent with tighter histone binding

on the left half due to the presence of highly flexible pyrimidine-purine dinucleotides at the inward-facing minor groove regions [51, 52]. The disparity in unwrapping behaviors also manifested as two distinct relationships between a pair of adjacent 6mA cluster and 6mA footprint (Fig. 3C). For a 6mA cluster on the left of a Widom 601 nucleosome, its size was largely independent from the size of the 6mA footprint for the nucleosome, indicating that the left 6mA cluster rarely encroaches into the nucleosome on its right. Their size distribution plot featured prominent horizontal and vertical stripes intersecting at the most probable states, roughly corresponding to sizes of the linker and nucleosomal DNA (Fig. 3C). For a 6mA cluster on the right of a Widom 601 nucleosome, its size mostly covaried in a negative linear relationship with the size of the 6mA footprint for the nucleosome (Fig. 3C). The diagonal stripes, also passing through the most probable states, indicate that the right 6mA cluster frequently encroaches into the nucleosome on its left. Thus, by analyzing both 6mA clusters and the complementary 6mA footprints, we can accurately determine nucleosome position and dynamics in the Widom 601 nucleosome array.

To apply 6mA-FP to study human chromatin, we used M.EcoGII, a prokaryotic MTase targeting adenine sites in any sequence context [53], which was able to deposit 6mA at higher density and generate more prominent footprints with less sequence bias than Hia5 (Additional file 1: Fig. S7) [14, 35, 54]. Nuclei isolated from human acute myeloid leukemia (AML) cell lines MOLM13 and OCI-AML3 were subjected to short (1 h) or long in vitro methylation (3 h) by M.EcoGII (Additional file 1: Fig. S7) [55, 56]. After in vitro methylation, genomic DNA was digested by DpnI (targeting fully methylated GATC sites) to generate fragments enriched for 6mA and of the preferred size range (3–5 kb) for accurate 6mA calling (Additional file 1: Fig. S7). The 6mA-FP samples were sequenced on both Sequel and Revio systems, and 6mA sites were called by ipdTrimming (Additional file 2: Table S4). We detected many clustered 6mA sites on individual DNA molecules (Fig. 3D), separated by a distance smaller than the linker DNA length ($78.8\% \leq 15$ bp, $84.1\% \leq 25$ bp, $86.8\% \leq 35$ bp) (Fig. 3E). We plotted the size distribution

(See figure on next page.)

Fig. 3 Nucleosome position and dynamics revealed by 6mA-FP. **A** 6mA distribution on the Widom 601 nucleosome array after 6mA-FP. Top: a schematic diagram for positioned nucleosomes and linker DNA at regular intervals. Middle: 6mA distribution in aggregate (shown as 6mA penetrance) and in individual DNA molecules (sorted by the 6mA count). Bottom: zoom-in on a DNA molecule, illustrating the definition of the 6mA clusters and footprints. **B** 6mA distribution on individual nucleosomes and their flanking linker DNA. Top: 6mA distribution in aggregate (shown as 6mA penetrance), superimposed on a schematic of a nucleosome positioned on the Widom 601 sequence; the nucleosome dyad, regions protected by H3-H4 tetramer, H2A-H2B dimer, H3 α N helix, as well as linker DNA are indicated. Bottom: 6mA sites on individual DNA molecules (sorted by the 6mA count). **C** Relationship between the size of adjacent 6mA clusters and footprints. For pairs with 6mA cluster on the left and nucleosome footprint on the right, horizontal, and vertical stripes are prominent. For pairs with 6mA cluster on the right and nucleosome footprint on the left, diagonal stripes are prominent. **D** Calling 6mA sites, 6mA clusters, and 6mA footprints on an individual human chromatin fiber after 6mA-FP. Top: IPDr values for all adenine sites on the DNA molecule. 6mA sites (red) and unmodified A sites (black) are separated by the IPDr threshold for 6mA calling (dashed line). Bottom: illustrating the definition of the 6mA clusters and footprints. **E** Distribution of the distance between adjacent 6mA sites on individual DNA molecules from short (blue) or long methylation (red) 6mA-FP of human chromatin. Most 6mA sites are close to each other (inset). Note major peaks in the long methylation (red) sample corresponding to the nucleosome ladder. **F** Relationship between the size of adjacent 6mA clusters and footprints from 6mA-FP of human chromatin. Compare short (blue) and long (red) methylation. Note the inverse relationship between the size of adjacent 6mA clusters and footprints (dashed line) on both short (blue) and long (red) methylation; a grid-like pattern with ~10 bp separation is only obvious for short methylation, corresponding to stepwise unwrapping of nucleosomal DNA

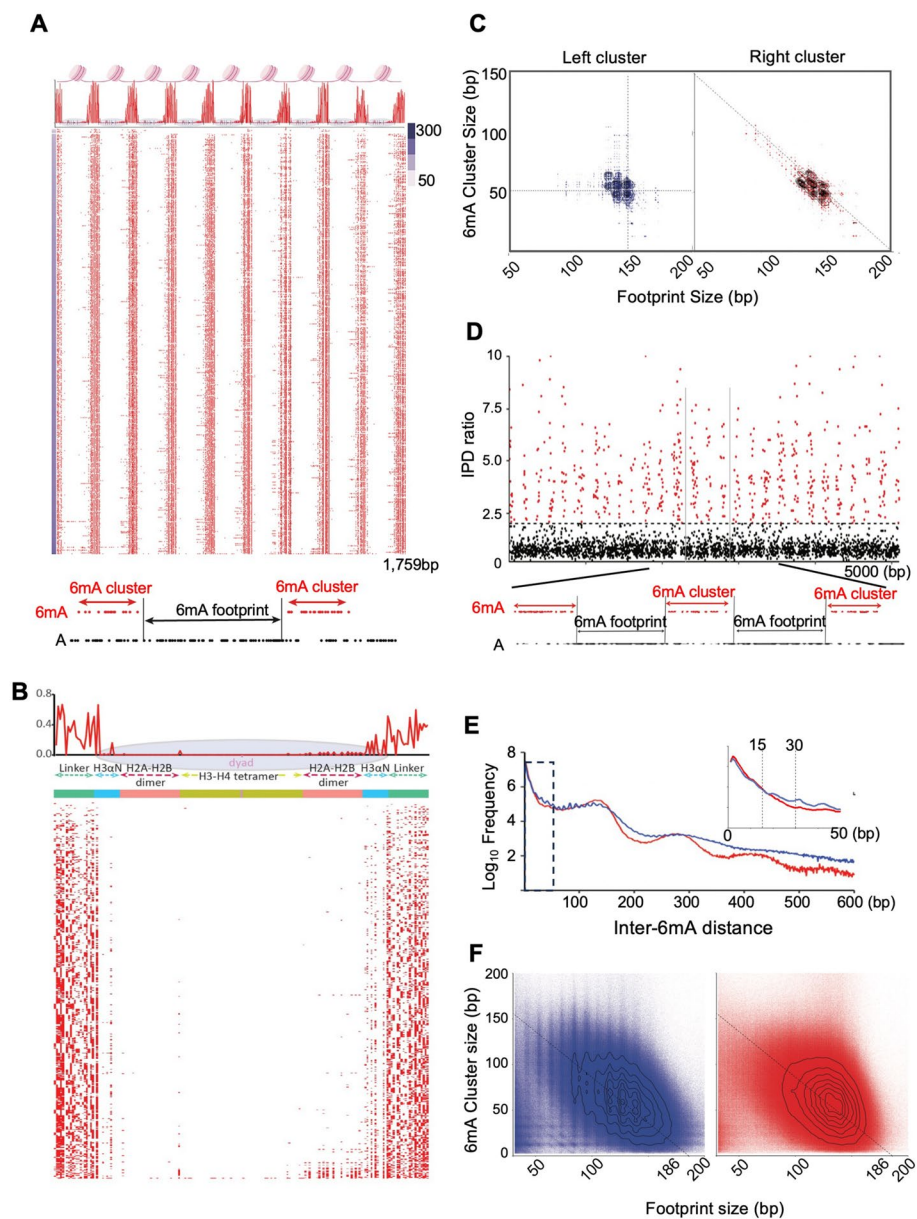


Fig. 3 (See legend on previous page.)

of all pairs of adjacent 6mA cluster and footprint in short or long methylation samples (Fig. 3F). For short 6mA-FP, the distribution showed a mixed pattern: (1) a diagonal feature was generated by 6mA clusters encroaching into nucleosomal 6mA footprints, (2) horizontal and vertical stripes were indicative of resistance to unwrapping of nucleosomal DNA, with a 10-bp grid consistent with stepwise unwrapping (Fig. 3F). For long 6mA-FP, the diagonal feature became more prominent but also more diffusive, while the horizontal/vertical feature was diminished, as resistance to unwrapping was overcome by nucleosome dynamics over time (Fig. 3F). These results are consistent with the model based on the Widom 601 nucleosome array (Fig. 3C) and validate that 6mA-FP can effectively reveal nucleosome position and dynamics in human chromatin.

Regularity and period of nucleosome arrangements in human cells revealed by 6mA-FP

Autocorrelation analysis of 6mA sites showed periodic distribution on most DNA molecules, with large oscillating periods of ~ 200 bp (Fig. 4A). This pattern is underpinned by the beads-on-a-string arrangement of nucleosomes constituting the 10-nm chromatin fiber, with the period for 6mA oscillation corresponding to the nucleosome repeat length (NRL) [57]. Nonetheless, 6mA oscillation exhibited substantial variations in its regularity and period among individual DNA molecules (Fig. 4A). We performed spectral clustering of auto-correlograms to distinguish five classes of DNA molecules (Fig. 4A–C, Additional file 1: Fig. S8) [58]. Regularity and period were the two principal determinants for the classification. Class I molecules showed low regularity and broad period in 6mA distribution (Fig. 4B, C, Additional file 1: Fig. S8). In contrast, Class II–V molecules had a more regular 6mA distribution and progressively longer 6mA period, likely corresponding to NRL of approximately 179, 189, 195, and 204 bp, respectively (Fig. 4B). We plotted these five classes of molecules for direct visualization based on periodicity and regularity together (Fig. 4C) or individually (Additional file 1: Fig. S8). Functional annotation showed that Class I molecules were enriched for promoters and enhancers of active genes but depleted for repetitive sequences (Fig. 4D). This is consistent with chromatin opening and nucleosome disruption at transcription regulatory elements, as opposed to the more closed heterochromatin regions with more regular nucleosome arrays [59].

It has long been assumed that α -satellite repeats constituting the centromeric heterochromatin can strongly position nucleosomes [60]. 6mA-FP allowed us to examine chromatin organization at human centromere, fully sequenced only recently [61], at the single molecule level. Centromere regions had regular nucleosome distribution, corresponding to Class II–V molecules (Additional file 1: Fig. S9). Unexpectedly, we found substantial heterogeneity in 6mA periods among DNA molecules mapped to the representative α -satellite higher-order repeat (HOR) J1/J2, despite essentially the same underlying sequence (Fig. 4E, F, Additional file 1: Fig. S9). This was reflected by the diversity in their classification by spectral clustering (Additional file 1: Fig. S9). Furthermore, nucleosome positioning was only observed near the ends of these DNA molecules, but it decayed very quickly inside (Fig. 4E). This result illustrates that 6mA-FP can reveal the intrinsic heterogeneity in chromatin organization at the single molecule level, which is often underestimated if only extrapolating from the ensemble level data.

Transcription factor (TF)-binding events and transcription-associated epigenetic landscape

Combinatorial binding of transcription factors (TFs) at promoters and enhancers controls transcription in eukaryotes [62–64]. TF usually resides in a nucleosome-depleted region (NDR) [64]. Its binding can lead to MTase protection, leaving behind a small footprint nested inside the NDR (Fig. 5A: left). However, TF binding is often very dynamic, and its MTase protection is readily overcome during 6mA-FP, leaving behind only a large 6mA cluster encompassing the NDR (Fig. 5A: right). To examine whether our optimized 6mA-FP procedure can detect dynamic TF binding events at the single molecule level, we first focused on CTCF, for the following reasons: (1) CTCF recognizes a long sequence motif and has potentially a larger footprint than most TFs [65–69]; (2) stably bound CTCF, likely in association with the cohesin complex [70], has longer residence

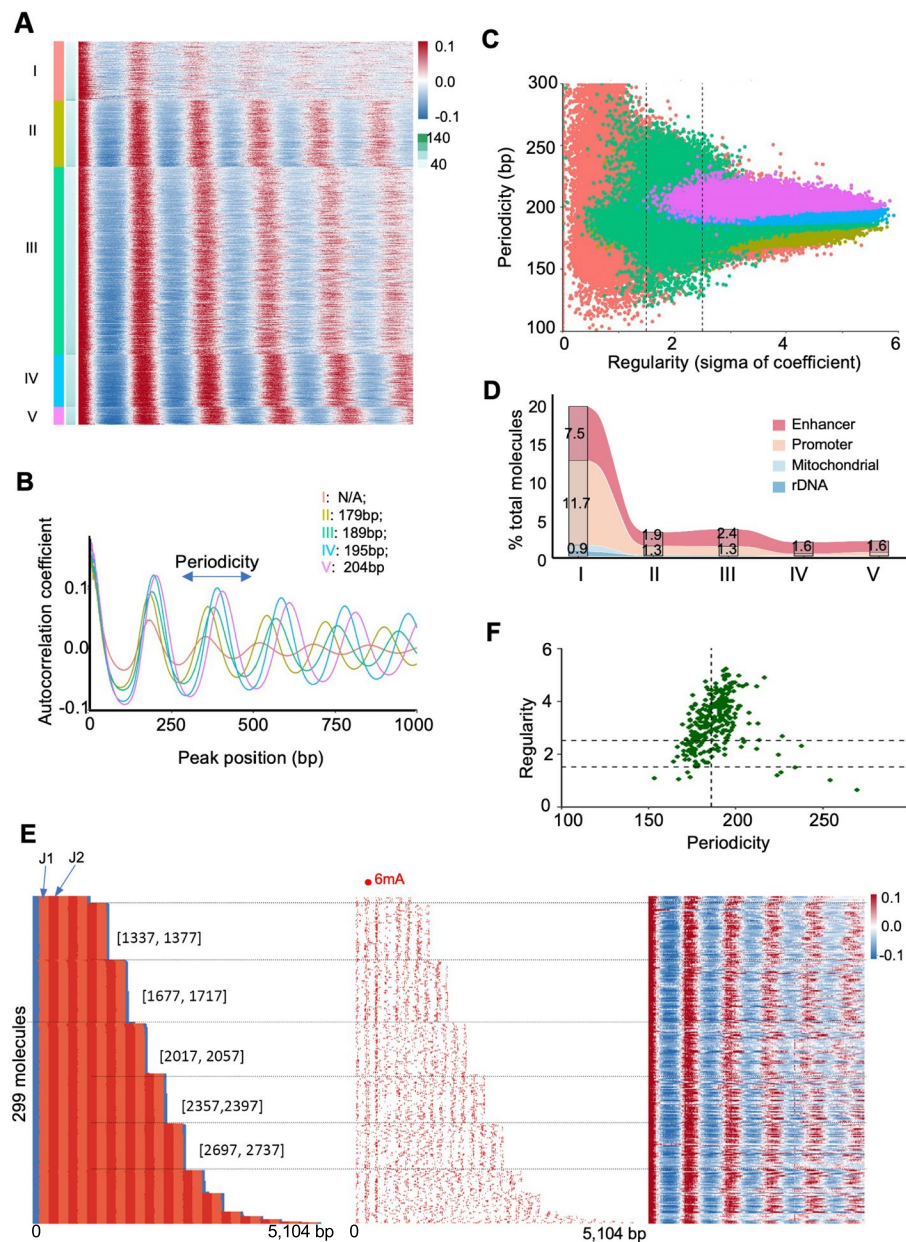


Fig. 4 Heterogeneity in human chromatin revealed by 6mA-FP. **A** Auto-correlograms of 10,000 randomly sampled DNA molecules, classified by spectral clustering (Class I–V) and ranked by 6mA counts. Top color bar: autocorrelation coefficient σ ; bottom color bar: 6mA density (count/kb). **B** Aggregated correlograms for each class. Nucleosome repeat length (NRL) estimation is shown for cluster II–V. **C** Scatter plot for direct visualization of single molecules from color-coded clusters shown in **A**, based on regularity (sigma of autocorrelation coefficient σ) and period (estimated NRLs). **D** Alluvial plot showing the proportional shifts of annotated molecules across each cluster. **E** 6mA signals at J1/J2 centromeric higher-order repeat (HOR). DNA molecules of J1/J2 HOR (color-coded for J1/J2 monomer, left), 6mA sites on individual DNA molecules (middle), and autocorrelation between 6mA sites (right) from 6mA-FP of human chromatin, ranked by read length. **F** Scatter plot showing the relationship between 6mA period (x-axis) and regularity (y-axis) in J1/J2 reads. Vertical dash line indicates the average NRL of 186bp

time than most TFs (in seconds) [69, 71, 72]; and (3) high-quality ChIP-seq data for CTCF are available in the human cell lines we studied (MOLM13 and OCI-AML3) [73–75]. To distinguish strong and weak CTCF binding sites, we ranked all genomic positions

with a high-confidence CTCF binding motif by their ChIP-seq signals (Fig. 5B, E, Additional file 1: Fig. S10). DNA molecules covering strong CTCF binding sites generally featured a small 6mA footprint, <50 bp in size, centered around the CTCF consensus sequence. These CTCF footprints were 5mC-hypomethylated, demarcated by densely methylated 6mA clusters, and often flanked by large 6mA footprints corresponding to well-positioned nucleosomes (Fig. 5B–D, Additional file 1: Fig. S10). In contrast, DNA molecules covering weak CTCF binding sites showed few small 6mA footprints centered around the CTCF consensus sequence, which were generally 5mC-hypermethylated (Fig. 5B–D, Additional file 1: Fig. S10). Moreover, nucleosome-sized 6mA footprints were randomly distributed across the weak CTCF binding sites, without any strong positioning (Fig. 5B–D, Additional file 1: Fig. S10). Prolonged treatment by M.EcoGII largely eliminated CTCF footprints at the strong binding sites and merged adjacent 6mA clusters into a large one, while few extended 6mA clusters were detected in DNA molecules covering the weak CTCF binding sites (Fig. 5E–G, Additional file 1: Fig. S10). DNA molecules with strong and weak CTCF binding exhibited clear visual contrast both individually and in aggregate (Fig. 5B–G). We conclude that the dynamic CTCF binding events at the single molecule level are sensitive to 6mA-FP, revealing distinct 6mA patterns at binding sites accompanied by varied phasing of flanking nucleosomes.

We extended our analyses to other transcription factors (TFs) highly expressed and functionally important in the human AML cell lines OCI-AML3 and MOLM13, including SPI1/PU.1, CEBP α , MEF2 C, MEF2D, IRF8, and MEIS1. By examining their ChIP-seq signals in the two cell lines, we found that these TFs were mostly enriched in and around large 6mA clusters (≥ 100 bp), but not the prevalent small 6mA clusters (50–60 bp) (Fig. 6A, B, Additional file 1: Fig. S11). Furthermore, only large 6mA clusters were preferentially flanked by active histone marks associated with promoters and enhancers, including histone H3K4 methylation and multiple histone acetylation events (Additional

(See figure on next page.)

Fig. 5 CTCF-binding events revealed by 6mA-FP. **A** Schematic for detecting TF binding by 6mA-FP. Small 6mA clusters indicate linker DNA and chromatin regions inaccessible to TFs; small 6mA footprints nested inside NDR indicate the TF-on state of chromatin regions accessible to TFs; large 6mA clusters encompassing NDR indicate the TF-off state of chromatin regions accessible to TFs. **B** CTCF binding detected by short methylation 6mA-FP. 6mA sites on individual DNA molecules centered around the CTCF binding motif are shown on the bottom. They are ranked from highest to lowest by CTCF ChIP-seq read counts (for a 118-bp window centered at the CTCF binding motif). Aggregated 6mA distribution around the CTCF binding motif are shown on top. **C** Small 6mA footprints revealed by 6mA-FP from short methylation reaction. DNA molecules are ranked as in Fig. 5B. 6mA footprints are color coded according to their size. **D** Heatmap for distribution of size (y-axis) and position relative to the CTCF binding motif (x-axis) of 6mA footprints in CTCF-high (Rank 1) and CTCF-low (Rank 5) DNA molecules. Most footprints are between 100 to 200 bp, corresponding to nucleosomes. Strong enrichment of small 6mA footprints (≤ 50 bp), representing the CTCF footprints, is found in CTCF-high but not CTCF-low molecules. **E** CTCF binding detected by long methylation 6mA-FP. 6mA sites on individual DNA molecules centered around the CTCF binding motif are shown on the bottom. They are ranked from highest to lowest by CTCF ChIP-seq read counts (for a 118 bp window centered at the CTCF binding motif). Aggregated 6mA distribution around the CTCF binding motif are shown on top. **F** Large 6mA clusters revealed by long methylation 6mA-FP. DNA molecules are ranked as in Fig. 5E. 6mA clusters are color coded by the size, shown at bottom. **G** Heatmaps showing the distribution of 6mA cluster size (y-axis) and position relative to the CTCF binding motif (x-axis) across CTCF-high (Rank 1) and CTCF-low (Rank 5) DNA molecules. Small 6mA clusters (30–100 bp) likely correspond to linker DNA. Strong enrichment of large 6mA clusters centered at the CTCF motif (≥ 100 bp), representing nucleosome-depleted regions (NDRs), is found in CTCF-high but not CTCF-low molecules

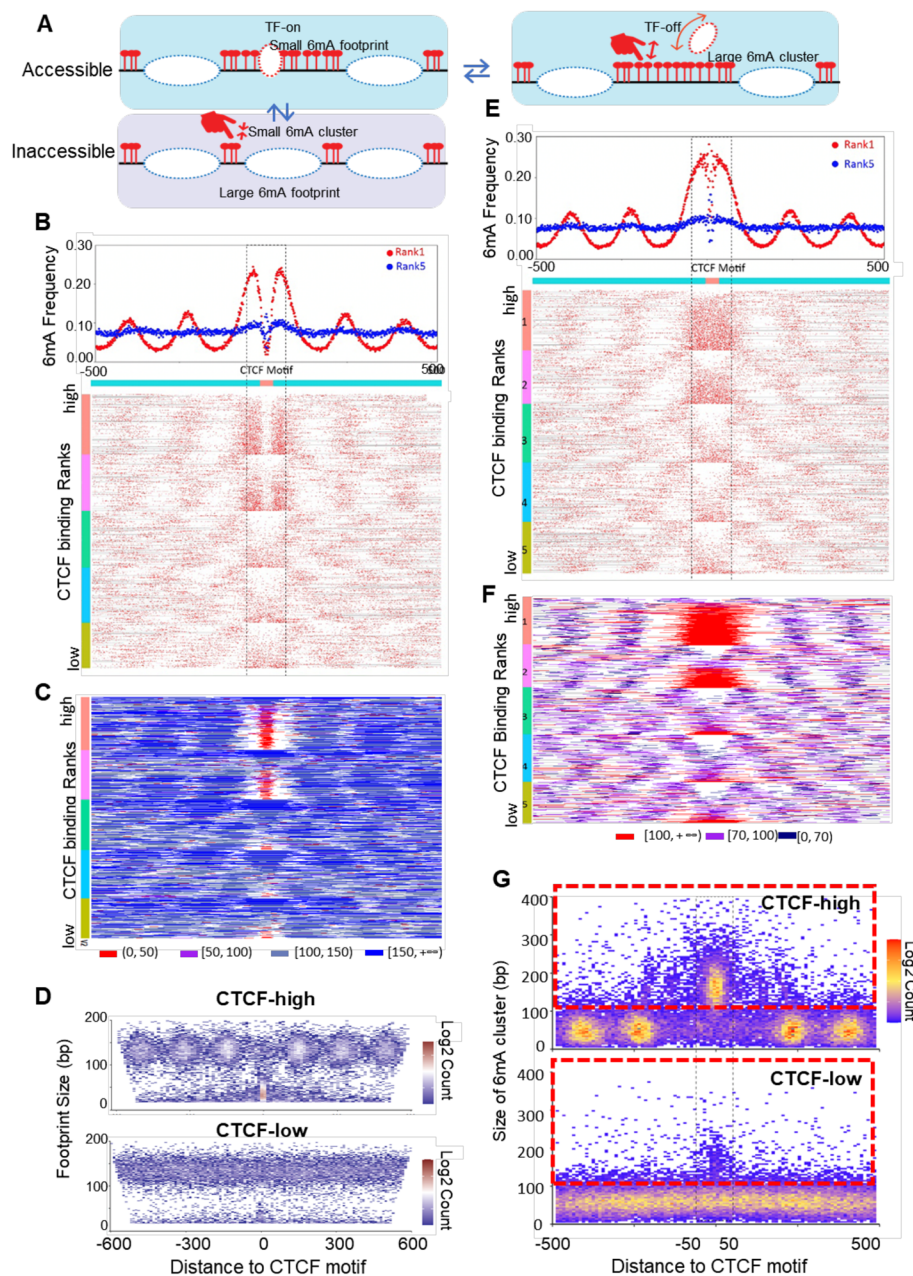


Fig. 5 (See legend on previous page.)

file 1: Fig. S11). In contrast, no preference for repressive histone marks was found near large 6mA clusters (Additional file 1: Fig. S11). To reveal heterogeneity in TF binding at the single molecule level, we examined SPI1/PU.1 and CEBP α binding on 28S rDNAs, which was deeply covered by SMRT CCS due to high copy numbers (Fig. 6C). Strong SPI1/PU.1 and CEBP α binding in the intergenic spacer region of 28S rDNA has been reported in the leukemia cells, with important roles in modulating nucleolar transcription and cell proliferation [76]. Here we found that for rDNA molecules covering the same genomic locus, only a fraction of them contained an overlapping large 6mA cluster, which disrupted the regular 6mA pattern associated with nucleosome arrays (Fig. 6C–E,

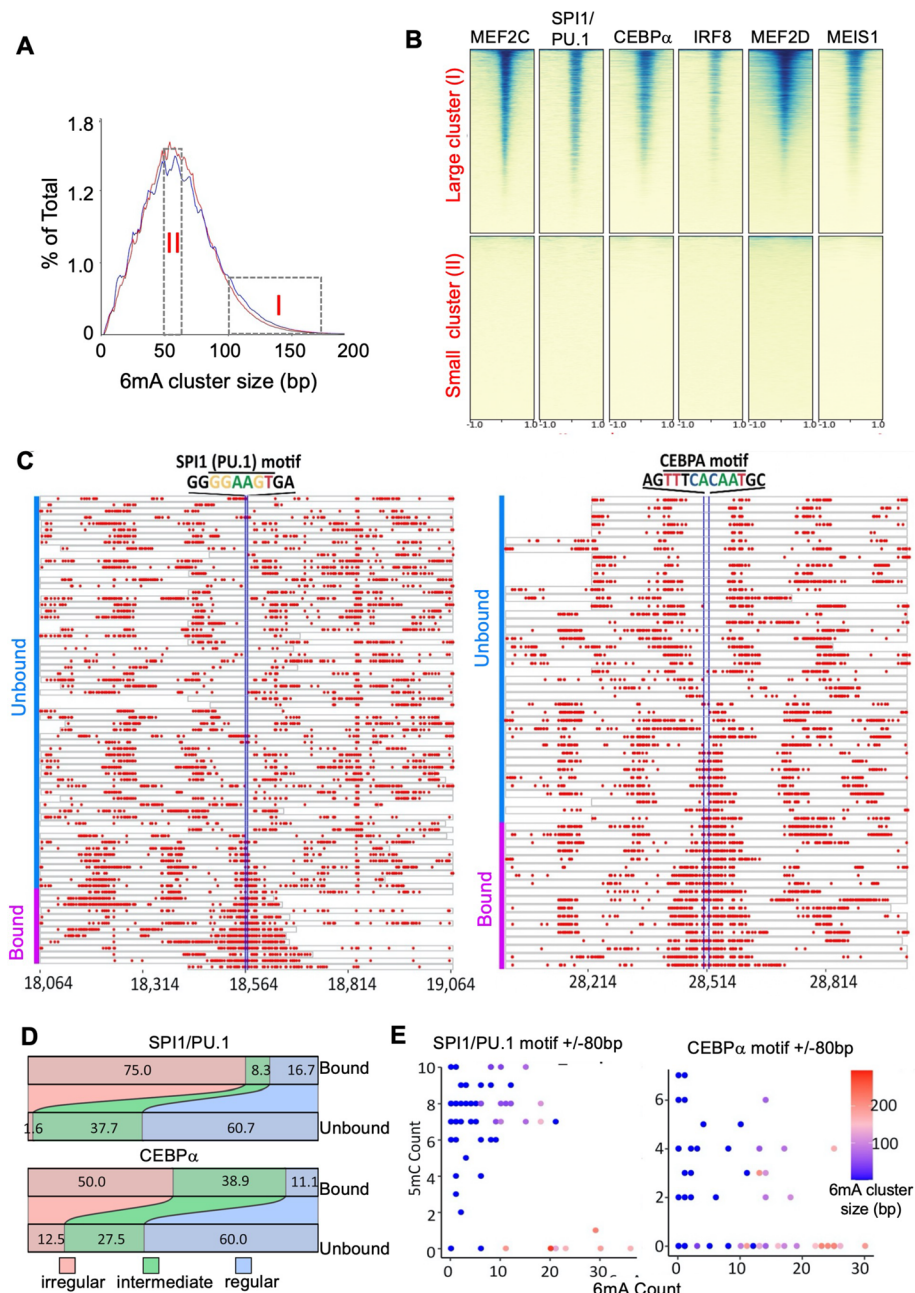


Fig. 6 TF binding at large 6mA clusters. **A** Size distribution of 6mA clusters. Two classes of 6mA clusters are defined according to their sizes: Class I (≥ 100 bp) and Class II ($60 \text{ bp} \geq \text{size} \geq 50$ bp). **B** Heatmap showing differential transcription factor binding (based on ChIP-seq data) in large (I) and small (II) 6mA clusters. **C** Representative binding of SPI1/PU.1 (left) and CEBP α (right) at a single locus in human rDNA on individual DNA molecules. DNA molecules spanning the respective binding motifs are sorted by the size of the 6mA cluster. Transcription factor binding states are indicated on the left. **D** Alluvial plot illustrating the regularity of 6mA patterns surrounding SPI1/PU.1 or CEBP α binding sites, comparing bound and unbound states. **E** Scatter plot showing the inverse relationship between 6mA and 5mC levels around SPI1/PU.1 and CEBP α consensus sequences. Color bar indicates the 6mA cluster size

Additional file 1: Fig. S12). In contrast, the rest of the rDNA molecules contained no (large) 6mA clusters at the binding motif and showed predominantly 6mA patterns indicative of regularly spaced—but randomly positioned—nucleosomes (Fig. 6C–E). An

inverse relationship between exogenous 6mA and endogenous 5mC was also detected for SPI1/PU.1 and CEBP α binding loci at the single molecule level (Fig. 6D, Additional file 1: Fig. S12). As the control, few rDNA molecules had large 6mA clusters at a CTCF motif although reciprocal patterns of 6mA and 5mC were observed (Additional file 1: Fig. S12). Our result is consistent with SPI1/PU.1 and CEBP α binding at selective rDNA molecules and subsequent recruitment of ATP-dependent chromatin remodeling and 5mC demethylation activities [77–80].

Regulatory signals for RNA polymerase II (Pol II) transcription are integrated through recruitment of transcription factors/cofactors, chromatin modifying/remodeling activities, and transcriptional/co-transcriptional machineries, generating specific transcriptional output in coordination with the epigenetic landscape [81]. We closely examined 6mA-FP data related to transcription-associated epigenetic changes. We focused on transcriptional start sites (TSS) of Pol II-transcribed genes, which were ranked and classified by their epigenetic or transcriptional states (Fig. 7). 6mA-FP revealed unusual patterns in 6mA clusters and footprints associated with genes enriched for two active histone marks—H3K4me3 and H3K27ac (Fig. 7A). While there was significant heterogeneity, a large fraction of DNA molecules contained large 6mA clusters and/or small 6mA footprints around TSS, which was obvious at both the single molecule level and in aggregation (Fig. 7B, C). Regularity of nucleosome arrangements was also disrupted on DNA molecules mapped to these TSS, which predominantly belonged to Class I in spectral clustering (Additional file 1: Fig. S13). In contrast, a much smaller fraction of DNA molecules exhibited a similar 6mA pattern for TSS that had either no H3K27ac or depleted for both marks, and they showed irregular 6mA distribution (Fig. 7A–C). We also ranked TSS by their transcript levels and found that a similar bias for large 6mA clusters and/or small 6mA footprints to be associated with TSS of highly expressed genes (Fig. 7D–F). These large 6mA clusters, hypomethylated at 5mC, were attributed to NDR generated by chromatin opening at core/proximal promoters (Additional file 1: Fig. S13). The notable small 6mA footprints embedded in the NDRs may reflect binding of TFs or stably bound Pol II, especially in the preinitiation complex (PIC) and promoter proximal pausing complex (PPP), as recently discussed in a fiber-seq analysis of transcription in *Drosophila* [82].

Nucleoid organization of mitochondrial DNA (mtDNA)

The small circular mitochondrial DNA (mtDNA: 16.5 kb; Fig. 8A) is present at hundreds to thousands of copies in human cells [83]. Nuclear-embedded mitochondrial DNA (NUMT), some resulting from recent and ongoing transfer of mtDNA into the nucleus, complicates the analysis of heteroplasmic mtDNA [84, 85]. Here we show that 6mA-FP can effectively distinguish NUMT and mtDNA by detecting their unique genetic and epigenetic signatures (Fig. 8B). Consistent with its nucleosome-free nature, mtDNA was densely decorated with 6mA, with very few nucleosome-sized 6mA footprints (Fig. 8C). In contrast, NUMT segments featured the periodic 6mA pattern indicative of the nucleosome-based chromatin organization (Fig. 8C). NUMT segments were further distinguished by high number of 5mC calls and mismatches, consistent with their nuclear presence and high mutation rates due to lack of selection (Fig. 8C, inset).

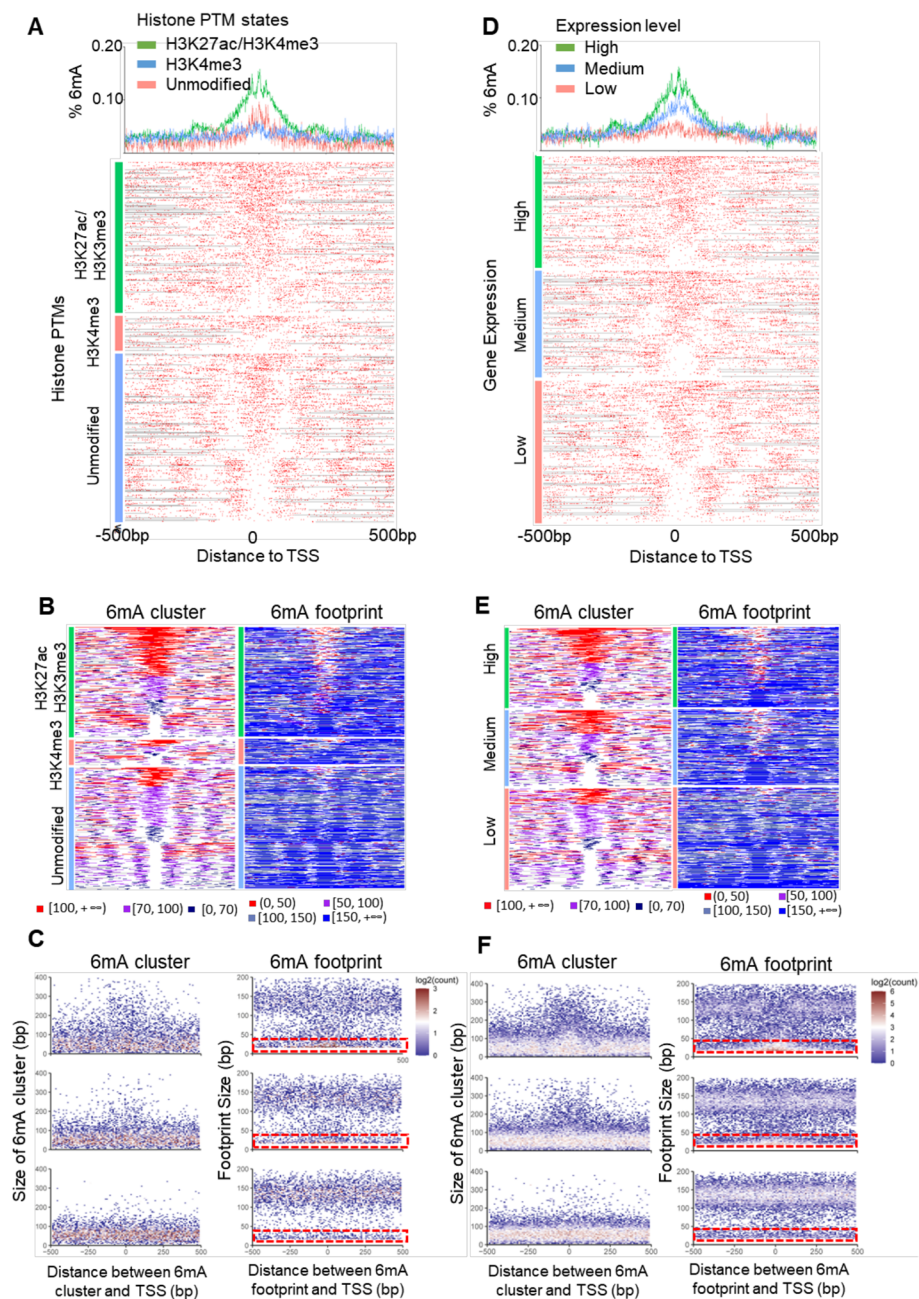


Fig. 7 Altered chromatin organization around transcription start sites (TSS). **A** 6mA distribution around TSS regions enriched for H3K27ac and H3K4me3, only H3K4me3, or neither mark. Top: aggregated 6mA signals, centered around the three categories of TSS. Bottom: 6mA sites on individual DNA molecules, grouped by epigenetic marks as indicated on left. The molecules in each group are sorted by distance of 6mA cluster to TSS. **B** Distribution of 6mA clusters (left) and footprints (right) on individual molecules. The molecules are grouped as indicated on left. The color indicated different sizes as indicated at bottom. **C** Heatmaps showing the distribution of 6mA cluster size (y-axis) and position relative to the TSS (x-axis) across DNA molecules in each category—clusters (left) and footprints (right). Dashed boxes highlight small footprints (≤ 50 bp). **D** 6mA distribution around TSS of genes with high, medium, or low expression levels. Top: aggregated 6mA signals, centered around the three categories of TSS. Bottom: individual DNA molecules showing 6mA sites, grouped expression level and sorted by the size of the 6mA cluster over TSS. **E** Distribution of 6mA clusters (left) and footprints (right) on individual molecules, classified and sorted as in Fig. 7D. Both 6mA clusters and footprints are color coded according to their sizes. **F** Heatmaps showing the distribution of 6mA cluster size (y-axis) and position relative to the TSS (x-axis) for each category—clusters (left) and footprints (right). Dashed boxes indicate small footprints (≤ 50 bp)

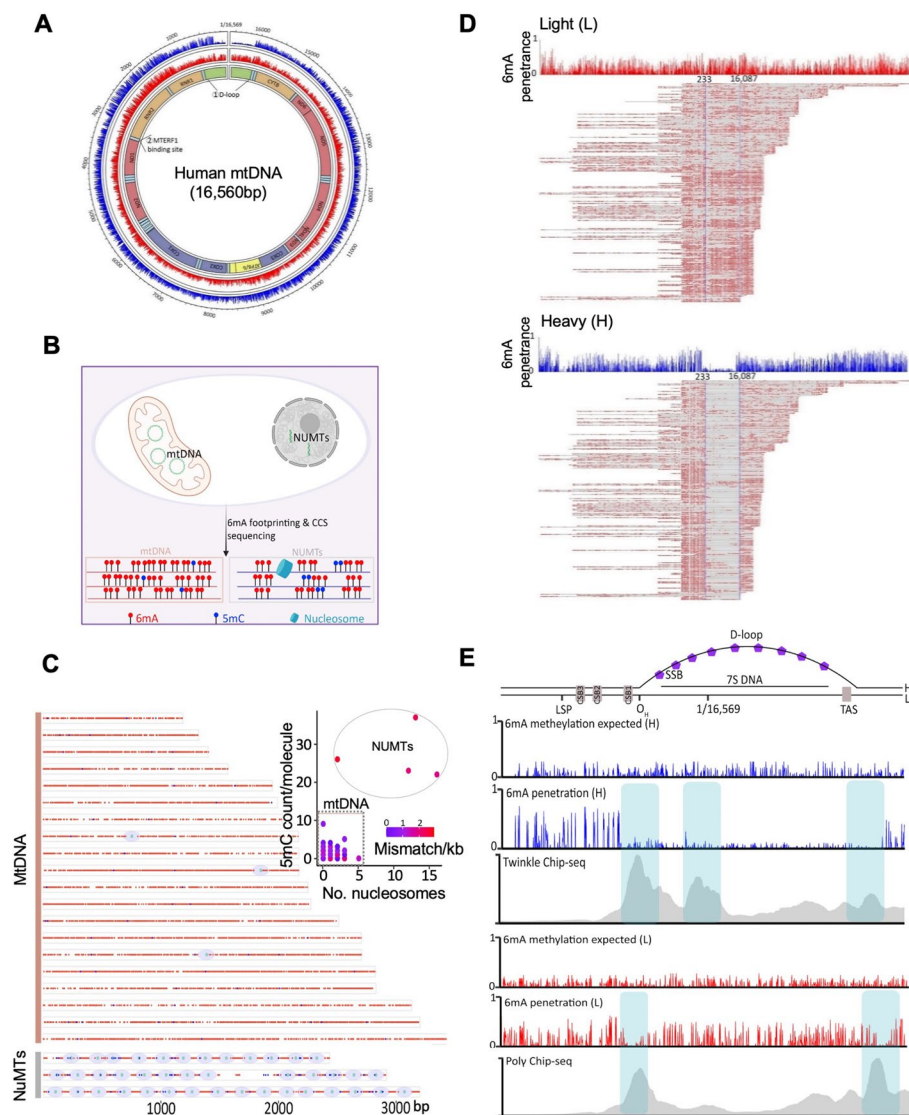


Fig. 8 Nucleoid organization of mitochondrial DNA (mtDNA). **A** Aggregated distribution of 6mA on human mtDNA. mtDNA annotations are shown in the inner circle, with 6mA distributions plotted separately for the heavy strand (blue) and light strand (red). **B** Schematics illustrating the features detected by 6mA-FP that distinguish between mtDNA and NUMTs. **C** Representative mtDNA and NUMT DNA molecules detected by 6mA-FP. Purple oval, a nucleosome called by the presence of a large 6mA footprint; blue dot, 5mC; red dot, 6mA. Inset, the scatter plot showing the number of detected nucleosomes vs. 5mC on individual molecules. Color scale indicates the number of mismatches to mitochondrial sequence. **D** 6mA distributions on the light (red) and heavy strand (blue) of the D-loop. Top, aggregated 6mA distribution. Bottom, individual DNA molecules aligned on the right to reference sequence. **E** The 6mA profiles for the heavy strand (blue) and light strand (red) of D-loop. Top, schematic for the D-loop and nearby genomic landmarks. Line plots show the expected 6mA methylation distribution based on methylation probability and the observed 6mA level. ChIP-seq tracks for Poly and Twinkle are included and their alignments to respective 6mA footprints are highlighted in blue shades

Human mtDNA is packaged into nucleoids with various mitochondria-specific nucleoproteins, especially TFAM [86]. In our 6mA-FP experiments with human cells, we recovered numerous mtDNA reads. M.EcoGII was more effective at methylating mtDNA than Hia5, even after short incubation (Additional file 1: Fig. S14).

With M.EcoGII, we found high resolution footprints for the mitochondrial transcription termination factor MTERF1 in most DNA molecules mapped to its binding site (Additional file 1: Fig. S14). Intriguingly, while 6mA levels from the coding region were similar for DNA molecules on both heavy and light strands, strong methylation bias in favor of the light strands were found on the displacement loop (D-loop) within the noncoding region of mtDNA (Fig. 8D and Additional file 1: Fig. S14). The D-loop is ~ 650 bp in length, duplexing with the light strand (double-stranded DNA/dsDNA) and displacing the heavy strand (single-stranded DNA/ssDNA) (Fig. 8E). The bias is most likely attributable to protection of the displaced heavy strand by the mitochondrial ssDNA binding protein (mtSSB) [53, 86, 87]. We noticed that the D-loop was flanked on both sides by sharp boundaries of very low 6mA densities. These 6mA footprints, present on both the heavy and light strand, aligned with ChIP-seq peaks of the mitochondrial-specific DNA helicase TWINKLE and DNA polymerase γ (POL γ) (Fig. 8E). TWINKLE is stably associated with the heavy strand [88, 89]; its footprints on the heavy strand were sharply bounded on one side by high 6mA density regions outside the D-loop and more loosely defined on the other by low 6mA density regions inside the D-loop (Fig. 8E). POL γ is stably associated with the light strand [90]; its footprints on the light strand were sharply bounded on both sides by high 6mA density regions within and without the D-loop (Fig. 8E). Importantly, two sets of POL γ footprints were unambiguously identified on both sides of the D-loop for almost all DNA molecules. These results strongly support that two sets of the mitochondrial replication machinery are loaded onto the D-loop, and they are stably arrested at the heavy strand replication origin (O_H) and the termination-associated sequence (TAS), respectively. This conclusion can only be rigorously derived from our single-molecule long-read 6mA-FP data, but not from the previously published ensemble level short-read ChIP-seq data. Our results not only demonstrate the ability of 6mA-FP in resolving large non-canonical DNA structures but also its power in establishing coordinate binding events at the single molecule level.

Discussion

Calling base modifications with SMRT CCS

As a sequencing by synthesis technology, SMRT/PacBio sequencing takes full advantage of DNA polymerase Φ 29's high fidelity, robust strand displacement activity, and strong processivity to generate multiple strand-specific readouts (subreads) of the same DNA molecule. Facilitated by sequence complementarity of DNA double helix, SMRT sequencing can generate high-quality primary sequence for a DNA molecule with only a few Φ 29 passes (for regular HiFi sequencing: ≥ 5 passes for each strand, equivalent to ≥ 10 passes for each DNA molecule). However, it is more challenging to call base modifications, which are distinguished from their unmodified counterparts by polymerase kinetics and can be present in a strand-specific manner. SMRT CCS detects 6mA by registering its kinetic signature of significantly increased IPD, relative to an unmodified A of the same sequence context (i.e., increased IPDr). However, multiple Φ 29 passes are needed for precise evaluation of IPD, as each subread IPD readout represents stochastic sampling of IPD values distributed in an exponential decay [42]. Kinetic analysis is further complicated by occasional prolonged polymerase pausing. Reducing the number of

Φ 29 passes dramatically increases the probability of overestimating IPD and miscalling 6mA. We therefore set a high threshold (≥ 10 passes for each strand, equivalent to ≥ 20 passes for each DNA molecule) for 6mA calling and optimize DNA fragment sizes (3–5 kb) in SMRT CCS library preparation accordingly. This high threshold also allows us to add a local outlier trimming step in IPD conversion to counter the effect of polymerase pausing, by removing the maximum subread IPD value for each site before averaging the rest to output a CCS IPD value (equivalent to removing $\leq 10\%$ of subread IPD values, given the requirement for ≥ 10 passes for each site). This IPD conversion step, implemented by a modified CCS module, is critical for our ipdTrimming pipeline and consistent with the long-standing practice of outlier removal [20, 44]. As a trade-off, setting a high threshold of Φ 29 passes undoubtedly reduces SMRT CCS sequencing coverage, which is directly limited by the smaller DNA fragments in the sequencing library (3–5 kb for 6mA detection vs. ~ 10 kb for regular HiFi sequencing).

Current 6mA detection methods using SMRT CCS data fall into two major categories: (1) statistical modeling-based, such as ipdSummary [34, 39], SAMOSA [16], and 6mA-sniper [91], which use techniques like Gaussian Mixture Models (GMMs) to detect the 6mA kinetic signature in the subread data, but incur high computation cost and reduced scalability; (2) machine learning-based, such as Fibertools [35] and Jasmine [36], which leverage neural network to predict 6mA with the CCS data, but face challenges in obtaining comprehensive training datasets and recapitulating critical kinetic information. We emphasize that if the input data quality is poor—as is the case for the subread data with a low number of Φ 29 passes, or the CCS data with crosstalk from polymerase pausing—no amount of processing can lead to reliable 6mA calls. Our analysis has demonstrated that the 6mA kinetic signature in the subread data can be efficiently recapitulated by ipdTrimming in the CCS data, which in turn can be used as input for simple Gaussian demixing procedures or sophisticated machine learning algorithms to make accurate 6mA calls.

Our analysis is mainly focused on 6mA's primary onsite effect on IPD, directly attributable to the kinetic barrier for the incoming dTTP to pair with 6mA on the template DNA strand. Nonetheless, 6mA also has secondary offsite effects, as the template DNA strand makes extensive contact with DNA polymerase Φ 29 and the N^6 -methyl group may indirectly affect IPD at nearby sites [46, 92]. With high 6mA density, as is the case in 6mA-FP, its secondary offsite effects are likely to increase [20]. 6mA's secondary offsite effects are most obvious at positions 5 nucleotides upstream (5') of primary 6mA sites [20, 46]. Increased IPD values at secondary sites, if an adenine, may cause their misidentification as 6mA. Intriguingly, ipdTrimming makes substantially fewer 6mA calls at one of the likely secondary sites than ipdSummary. The local trimming of IPD outliers may therefore also limit 6mA's secondary offsite effects, contributing to ipdTrimming's better performance with 6mA-FP data. Crosstalk between nearby base modifications (6mA and 5mC), mediated by their secondary offsite effects, complicate their calling. In this study, we use native genomic DNA instead of WGA as a negative control for 6mA-FP of human chromatin, so that the kinetics baseline will include perturbations to IPD caused by the endogenous 5mC. High-quality ground truth datasets, with synthetically defined base modifications and their combinations in a wide range of sequence context, are needed to fully resolve this difficult issue.

Revealing epigenetic heterogeneity with 6mA-FP

In 6mA-FP, information about nucleoprotein organization is transcribed via in vitro methylation into exogenous 6mA on individual DNA molecules, which can be simultaneously revealed with endogenous 5mC signifying the epigenetic state of DNA, as well as the primary DNA sequence [18]. Importantly, each binding event can be registered by multiple methylation sites at the single molecule level, allowing it to be called with high confidence. Here we show that by finetuning the in vitro methylation timing, we can differentiate the TF-unbound state, the TF-bound state, and the TF-transient-off state of chromatin. Short methylation, resulting in low, unsaturated 6mA levels, can preserve footprints left by dynamic binding of CTCF and other chromatin-associated proteins. Long methylation, on the other hand, resulting in near saturation 6mA levels, can reveal the dynamic nature of these binding events and distinguish them from stable binding events as well as the unbound state. Also, by optimizing the in vitro methylation condition and enriching heavily methylated DNA molecules of proper sizes, our work generates accurate and high-resolution footprints that allow for unambiguous detection of the mtDNA D-loop and the strand-specific and coordinate loading of the mitochondrial replication machinery. This represents a great improvement from previous report of the nucleoid organization of human mtDNA [86]. It also shows that 6mA-FP can be used to resolve large non-canonical DNA structures and establish joint and strand-specific binding events at the single molecule level, both of which cannot be achieved by the ensemble level short-read ChIP-seq data.

6mA-FP can provide near-base pair resolution, much higher than what 5mC-based NOME-seq can achieve: on average, an adenine site is present for every 2 bp, while a GpC dinucleotide, targeted by the methyltransferase M.CviPI, is present at a much lower density (> 20 bp) [13, 93, 94]. With our optimized experimental and computational framework, we have extensively mapped DNA binding events around TSS at the single-molecule level, revealing critical features of the transcription-associated epigenetic landscape at unprecedented resolution in human cells. Here, we define multiple closely spaced 6mA sites as a 6mA cluster, which is also referred to a methylase-sensitive patch (MSP) [18]; complementarily, the interval between adjacent 6mA clusters, if containing multiple unmodified A sites, is designated a 6mA footprint. While we focus on the size distribution of 6mA clusters and footprints, statistical and machine learning approaches can be used to improve their classification and assignment to specific molecular events [27, 82, 86]. Critically, small 6mA footprints (< 50 bp) are highly enriched near active TSS, while nucleosome footprints are depleted. As recently reported in a fiber-seq analysis of transcription in *Drosophila* [82], stably bound Pol II, especially the preinitiation complex (PIC) and promoter proximal pausing complex (PPP), may contribute to these 6mA footprints. Sub-nucleosome particles (tetramer and hexamer), either standalone or in association with Pol II or chromatin remodelers, represent another alternative scenario [95]. However, we would like to point out that it is still a challenge to distinguish various binding events at the single molecule level, due to overlapping signals, background noise, and lack of specific recognition. Deeper understanding of the transcription-associated chromatin landscape at the single molecule level, especially over the range spanning transcription regulatory elements and transcriptional units, will provide critical insights into transcriptional regulation.

SMRT CCS is ideally positioned to provide the long-range, multimodal readout of genetic and epigenetic information at the single molecule level. To ascertain the overall occupancy and dynamic behavior of a nucleoprotein, many DNA molecules covering the same binding site need to be examined. This demand for high sequencing depth is a challenge to meet for mammalian systems with large nuclear genomes. Only for the high copy number rDNA and mtDNA, the heterogeneity of bound and unbound DNA molecules is fully revealed, even though it is likely a general phenomenon across the entire genome. The situation may be improved by the greatly expanded capacity of the Revio system. Reduced representation, analogous to Reduced Representation Bisulfite Sequencing (RRBS) for 5mC profiling [33, 96], and target sequence enrichment, such as CRISPR-CATCH [97], promise to further increase coverage for genomic regions of interest and warrant future studies.

Conclusions

The ipdTrimming pipeline establishes a robust and efficient framework for in vivo 6mA detection, facilitating a wide range of epigenomic applications. By incorporating outlier trimming into kinetics-based 6mA profiling, ipdTrimming achieves high precision and computational efficiency across all PacBio sequencing platforms. Furthermore, its integration into an optimized 6mA footprinting (6mA-FP) strategy enables high-resolution mapping of nucleoprotein interactions, allowing the detection of transcription factor binding dynamics at the single-molecule level. It also supports the analysis of long-range DNA-binding events in mitochondrial DNA. Collectively, our work provides a transformative tool to leverage long-read single-molecule sequencing to gain insights on epigenetic events across diverse biological contexts.

Methods

SMRT CCS data processing

ipdSummary

SMRT CCS subread data were processed as previously described [34]. Briefly, each DNA molecule in the raw data was partitioned into a separate BAM file using pysam module [98]. Circular Consensus Sequence (CCS) was calculated for each DNA molecule using the CCS module [37]. Only CCS reads with high passes (≥ 10 passes for each strand) were retained for the following analysis. Single-molecule aligned BAM files were generated using BLASR [99], which in turn served as the input for the ipdSummary module to calculate IPD ratios (IPDr) [37]. DNA molecules with global or local anomalies in IPDr were removed to further improve 6mA calling accuracy.

ipdTrimming

Local trimming of subread IPD outliers was introduced into IPD conversion: for each site in a DNA molecule, the maximum of subread IPD values were removed; the remaining subread IPD values were then averaged to generate the CCS IPD value. This was implemented by a modified CCS module [100], which used the raw subread data as the input and generated the CCS IPD values aligned to its own CCS sequence as the output. This was followed by movie-time normalization: to compensate for variations

in the polymerase elongation rate on individual DNA molecules, each CCS IPD value was normalized against the CCS IPD value averaged across all sites in a DNA molecule. This CCS IPD value was then compared to that of an unmodified base with the same sequence context, provided by a kmer model pretrained on the Sequel data [45]. The last two steps were implemented by a custom script (`ipdRatiocalculator_fromCCS.py` [100]), which output a BAM file for DNA sequence tagged with kinetics parameters, including IPD ratios (IPDr).

Revio Data processing

The on-instrument IPD conversion for the Revio system only outputs the CCS level IPD, by averaging all subread IPD values for each site of a DNA molecule. We adapted a kmer-based pipeline for 6mA calling on the Revio system. Specifically, we trained a new kmer model using the Revio data. We began by processing two input datasets: Revio-generated SMRT CCS results from 6mA-FP of human chromatin (6mA-FP HiFi.bam) and the native human genomic DNA (gDNA HiFi.bam). After movie-time normalization, gDNA HiFi.bam reads were aligned to the human reference genome (hg19) using `pbbmm2`, generating reference IPD values for each genomic position. We then trained a new kmer model based on the Revio-sequenced human genome. Specifically, CCS IPD values across all reads covering a certain adenine in the genome (sequencing coverage depth $\geq 10 \times$) were averaged to assign a reference/baseline IPD value to this genomic position. Genomic positions sharing the same local sequence context ($[-10, +4]$: NNNNANNNNNNNNNN; the same as the kmer model pretrained on the Sequel data) were subsequently binned, and their IPD values were averaged to finetune the new kmer model. To calculate IPDr, IPD values from 6mA-FP were compared with reference IPD values of the equivalent genomic position or kmer sequence context. We made three sets of comparisons: (1) to the kmer model pretrained on the Sequel data, (2) to reference IPD values assigned to the human genome, and (3) to the new kmer model trained on the Sequel data.

6mA calling

We first set an IPDr threshold that recalled 98.5% of GATC sites in the *dam*⁺ plasmid DNA positive control dataset, and then applied this threshold in the WGA negative control dataset to calculate the background noise level for 6mA calling. An alternative approach, by comparing the *Tetrahymena* native genomic DNA and WGA datasets, led to the same conclusion that the background noise level is much higher for the kmer pipeline (Additional file 1: Fig. S2). In samples with only a fraction of adenine sites methylated, their IPDr values exhibited a bimodal distribution. After logarithmic transformation, the 6mA peak with higher IPDr values was closely fitted by a Gaussian distribution curve. Gaussian demixing was performed using a custom script (`m6A_gaussianmix.py`), which determined an IPDr threshold above which 6mA was called, as well as the false positive and false negative rates for 6mA calling.

5mC calling

5mC was called by PacBio's `primrose` software with default parameters [101]. Methylation probabilities for each CpG site within the single molecule, present in the ML and

MM tags of the bam file, were converted into the continuous probabilities (ranging from 0 to 1) using `extract5mCforprimrose.py`. CpG sites with methylation probabilities ≥ 0.9 were preserved for downstream analysis.

6mA cluster and 6mA footprint

6mA sites detected at the single-molecule level were grouped into 6mA clusters. Specifically, all 6mA sites within 15 or 30 bp of each other were combined, and multiple combined 6mA sites (≥ 3) formed a 6mA cluster, the size of which was the distance between the first and the last 6mA sites. Following 6mA cluster identification, 6mA footprints were determined by analyzing regions between adjacent 6mA clusters. These inter-cluster regions, if containing multiple unmodified A sites (≥ 3), formed a 6mA footprint, the size of which was the distance between adjacent 6mA clusters. Two alternative thresholds for combining adjacent 6mA sites into a 6mA cluster were used: the higher threshold (inter-6mA distance ≤ 30 bp) robustly detected large 6mA footprints corresponding to nucleosomes; the lower threshold (inter-6mA distance ≤ 15 bp) was more sensitive at detecting small 6mA footprints such as CTCF and Pol II, at the cost of higher background noise.

6mA autocorrelation analysis and classification by spectral clustering

For 6mA-FP analysis, we first filtered out DNA molecules with few 6mA sites (< 30 6mA sites per kb). For autocorrelation analysis of 6mA sites, a vector consisting of a series of 0's (no 6mA) and 1's (6mA on either strand) was constructed for each DNA molecule. This vector was processed by the `acf` function within the `statsmodel` [102] python package to generate the full-length auto-correlogram. The head 1000 bp of the auto-correlogram for each DNA molecule was used for spectral clustering in the `sklearn` package [58]. The number of clusters was manually adjusted to avoid generating large numbers of clusters containing few DNA molecules.

We evaluated the periodicity and regularity of 6mA distribution in each DNA molecule. Briefly, autocorrelation functions were computed for individual 6mA footprint molecules to quantify the spatial correlation of 6mA signals across the DNA fibers. Peaks in the autocorrelation function correspond to periodic features, such as the NRL, while the amplitude of these peaks reflects the regularity of nucleosome spacing. For each DNA molecule, a vector consisting of a series of 0's (unmodified A on either strand), NaN (C/G), and 1's (6mA) was constructed. A 33 bp NaN-sensitive rolling mean was applied to the raw vector, enhancing signal smoothness using the R `imputeTS` package [103]. Autocorrelation analysis was then performed on the smoothed vector. Peaks (maximum and minimum) were identified by the `find_peaks` function in the python `scipy` module [104]. The absolute σ values (autocorrelation coefficient) of the first 8 peaks were summed up as an indicator for the regularity of 6mA distribution. The average inter-peak distance, obtained by linear regression of peak positions, was an estimate of the period of 6mA distribution, corresponding to the nucleosome repeat length (NRL) on individual chromatin fibers.

Analysis of centromeric/peri-centromeric reads

CCS reads were aligned to the human T2T-CHM13v1.0 [105] reference genome using *blastn* [106] to selectively enrich the (peri)centromeric reads in silico. Using the latest annotation of the centromeric regions in the T2T genome [61], molecules fully contained within these regions were retained. For the centromeric higher-order repeat (HOR) reads, molecules were represented as a series of monomers as previously described [107]. Briefly, a catalog of distinct 58 monomers were mapped back to the long read with these following parameters: `-max_target_seqs 1,000,000 -word_size 7 -qcov_hsp_perc 60`. The assignment of monomer to a long read was calculated via dynamic programming to ensure that each segment of the read be assigned at most one monomer. All code used for these analyses is derived from the GitHub link [108]. Importantly, a conserved GATC site in J1 allowed us to enrich DNA molecules comprising J1/J2 tandem arrays after 6mA-FP and DpnI digestion. It probably also promotes nucleosome positioning near both termini of J1/J2-containing DNA fragments, but not internally.

Transcription factor binding analysis

Transcription factor binding motifs in the MEME format, containing a collection of position weight matrices, were downloaded from the JASPAR CORE database [109], in which a curated, non-redundant set of profiles were experimentally defined. These motif sequences were aligned to the reference genome using FIMO [110] of the MEME-Suite. Assessment of genomic binding associated with these motifs was measured by the FIMO score. Notably, a threshold of score ≥ 45 , corresponding to the p -value $\leq 3.17e - 05$, was employed to identify significant matches for subsequent analysis. In parallel, CCS reads were mapped back to the same reference sequence by *blastn*. Only CCS reads with a mapped coverage of $\geq 98\%$ and a mapped identity of $\geq 95\%$ were retained for further analysis.

BEDTools [111] was utilized to find DNA molecules overlapping with these transcription factor binding sites. TF binding sites were extended by 50 bp in both upstream and downstream directions. Reads that fully covered this extended region were extracted. By combing the coordinates of reads mapped to the genome, methylation sites on single molecule and the strand information of TF binding sites, normalized methylation profile was obtained by calculating the column mean within a window of around 100 bp, centered at each TF binding site.

Transcription start site metaplots

Transcription start sites used in this study were derived from the list of all known RefSeq transcripts in hg19 provided by HOMER v4.11.1 [112]. For each protein-coding gene, the TSS of this gene is determined with these following criteria: If multiple transcripts originating from the gene share a common TSS, the TSS with the highest occurrence frequency is designated as the TSS for this gene. If no such dominant TSS exists, the midpoint of the TSSs across multiple transcripts is chosen as the gene's TSS. The TSS for each protein-coding gene was then elongated by 50 bp in both upstream and downstream directions. TSS regions were classified based on

whether these regions overlap with a specific histone modification peak by at least 1 bp. In addition, the methylation profile around TSSs was calculated in an identical manner to that described above for TF sites. CCS reads were required to fully cover the extended TSS region. Finally, CCS reads were sorted by either the 6mA cluster size overlapping with these extended regions or the expression values of these TSS-associated genes for visualization.

Mapping CCS to mitochondrial DNA (mtDNA)

The complete circular human mitochondrial genome was downloaded from the NCBI GenBank database with the accession number NC_012920.1. To make the long reads spanning the start/end of the circular genome mappable, a linearized mitochondrial genome was generated by concatenating two tandem copies of human mtDNA. CCS reads were aligned to the modified reference using blast with the following parameters: -word_size 50, -outfmt 5. The XML format of blastn output was converted into a SAM file with the reference sequence by Blast2Bam [113]. Reads with $\leq 98\%$ mapping coverage or $\leq 95\%$ mapping identity were excluded. Since the two identical mitochondrial genome sequences were joined, in principle, reads that do not span the start/end region could be potentially aligned to two separate regions. To ensure unique mapping for each read, only reads aligned to the first one (with small reference coordinates) of these two regions were retained. In the SAM file, CIGAR values describe how the read aligns to the reference genome, including insertions, deletions, mismatches, and matches, allowing for the alignment at the single-nucleotide resolution. By parsing the CIGAR values in the SAM file, methylation sites on both strands of individual DNA molecules were mapped back to the mitochondrial genome.

Mapping CCS to ribosomal DNA (rDNA)

In the context of the hg19 genome assembly, a single repeat unit sequence (NCBI U13369.1) rDNA was inserted into chr13. To enhance the mappability of the long reads derived from the rDNA, a modified reference rDNA was created. This modified sequence consists of intergenic sequence (IGS), 45-s coding rDNA and IGS, effectively representing one and a half tandem copies of rDNA. CCS reads mapping was processed in an identical manner to that described for mtDNA. Briefly, reads were mapped using blastn and then converted into the SAM file. Low alignment quality reads ($\leq 98\%$ mapping coverage or $\leq 95\%$ mapping identity) were removed. Reads mappable to both IGS regions were allocated to the first.

Cell culture

Human acute myeloid leukemia (AML) cell lines OCI-AML3 (DSMZ no. ACC 582) and MOLM13 (DSMZ no. 554) were cultured in RPMI-1640 (GIBCO™ 11,875–119) supplemented with 10% fetal bovine serum (Sigma) and 1% penicillin/streptomycin (GIBCO™ 15,140–163). All cell lines were authenticated and tested for mycoplasma contamination using Southern Biotech Mycoplasma Detection Kit (13,100–01).

Sample DNA preparation

A *PvuII*-digested fragment of the pBluescript II SK(-) plasmid (~ 1 kb) purified from *dam*⁺ *E. coli* was used as the positive control sample for 6mA calling. Native genomic DNA was extracted from *Tetrahymena* WT cells (SB210, *Tetrahymena* Stock Center) using Wizard® Genomic DNA Purification Kit (Promega). REPLI-g Single Cell Kit (Qiagen) was used for whole genome amplification (WGA) to generate the negative control sample. DNA samples were sheared to 3–5 kb in length with Megaruptor (Diagenode Diagnostics) and prepared for SMRT CCS.

1.7×10^5 OCI-AML3 or MOLM13 cells were lysed in 0.5 mL of nuclear extraction buffer (20 mM HEPES pH 7.9, 10 mM KCl, 0.1% Triton X-100, 20% glycerol, 0.5 mM spermidine, $1 \times$ Protease Inhibitor Cocktail) for 8 min on ice. Nuclei were isolated by centrifuge at 600 *g* for 5 min with two washes of nuclear extraction buffer. Purified nuclei were methylated in a 30 μ L of 50 mM Tris-HCl pH 8.0, 2 mM EDTA, 0.5 mM EDGA, 160 μ M SAM (New England Biolabs), $1 \times$ Protease Inhibitor Cocktail, and 38 μ M recombinant M.EcoGII or Hia5 for 1 h (short methylation) or 3 h (long methylation) at 37 °C. Genomic DNA was extracted with Monarch® HMW DNA Extraction Kit for Cells & Blood (New England Biolabs), digested with *DpnI* overnight at 37 °C, and resolved on 1% agarose gel. DNA fragments between 3 and 5 kb were gel extracted using the QIAEX II Gel Extraction Kit (Qiagen). Extracted DNA was quantified with the Quant-iT™ dsDNA Assay Kits, high sensitivity (HS), and broad range (BR) (Invitrogen™ Q33120) on a Qubit™ 3 Fluorometer. Samples were submitted for SMRT CCS at the Genomics Core Facility at the Icahn School of Medicine at Mount Sinai.

Protein purification

Expression and purification of pA-M.EcoGII and pA-Hia5 were performed as previously described [114]. Briefly, bacterial culture was grown to 0.5 OD₆₀₀ in LB medium containing 50 μ g/mL kanamycin (Sigma) and induced overnight with 0.4 mM IPTG at 16 °C. Cells were lysed in high salt buffer (1 M NaCl, 10 mM Tris-HCl (pH 7.5), 10 mM Imidazole, 0.5% NP-40, 10% glycerol, 1 mM PMSE, and protease inhibitor cocktail), and sonicated. Proteins were purified on Ni-NTA beads and washed in low-salt buffer (250 mM NaCl, 10 mM Tris-HCl (pH 7.5), 20 mM Imidazole, 0.5% NP-40, 10% glycerol, 1 mM PMSE, and protease inhibitor cocktail) and eluted (250 mM Imidazole). Eluate was dialyzed overnight with ULP1, and the cleaved tag was removed by Ni-NTA binding.

Supplementary informationAdditional file1: Supplementary Figures S1-S14.

All supplementary figures in this study.

Additional file2: Supplementary Tables S1-S5. Table S1: The SMRT CCS datasets in this study. Table S2: Impact of strand-specific subread count cutoffs on false negative rate (FNR) and false positive rate (FPR) of 6mA calling by ipdTrimming. Table S3: Statistics of Gaussian de-mixing for 6mA and unmodified A peaks. Table S4: 6mA calling statistics for 6 mA-FP. Table S5: Comparing 6mA recall and false positive rates for different computation tools.

Peer review information

Tatiana Kutateladze and Wenjing She were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer review history is available in the online version of this article.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-025-03592-9>.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

We would like to thank Dr. Aaron M. Wenger from PacBio for providing the modified CCS module, insightful discussion, and critical reading of the manuscript. We would like to thank Dr. Sheng Li at University of Southern California for her critical reading and discussion of the manuscript.

Authors' contributions

W.T.Y. developed the computation pipeline, performed most analysis, and prepared the manuscript; X.Q.W and J.W. performed the 6mA-FP experiments, prepared samples for sequencing and contributed to data interpretation; F.W. helped with manuscript preparation; Y.L. and Y.D. conceptualized, designed the study, and wrote the manuscript. All authors contributed to writing and revising the manuscript. All authors read and approved the final manuscript.

Funding

This research was supported by National Science Foundation (MCB-2435178) to Y.L. and by Norris Comprehensive Cancer Center NCI P30 grant (P30 CA014089-45) and Jane Anne Nohl Division of Hematology and Center for the Study of Blood Diseases, University of Southern California to Y.D.

Data availability

The datasets generated and/or analyzed for the current study are available with access codes as described below. Raw sequencing data were deposited in the NCBI under accession numbers PRJNA1166413, which can be viewed through the following link: <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1166413> [115]. All code used in this study is available at <https://github.com/YifanLab/ipdTrimming> [100], under the AGPL-3.0 license. Data and code were deposited in Zenodo with doi: <https://doi.org/10.5281/zenodo.14602961> [116]. Previously reported PacBio sequencing data, including *Tetrahymena* native genome DNA (SRR16944761) [117], *Tetrahymena* whole genome amplification (SRR28602303) [118], fiber-seq in human OCI-AML3 (SRR17010848) [119], fiber-seq in Widom 601 (SRR13173521) [120], and a PvuII-digested fragment of the pBluescript II SK(-) plasmid (SRR27430828) [121] are publicly available.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with University of Southern California ethical standards.

Consent for publication

All authors consent to the publication of this manuscript.

Competing interests

The authors declare no competing interests.

Received: 30 September 2024 Accepted: 27 April 2025

Published online: 21 May 2025

References

- Rivera CM, Ren B. Mapping human epigenomes. *Cell*. 2013;155:39–55.
- Mehrmohamadi M, Sepehri MH, Nazer N, Norouzi MR. A comparative overview of epigenomic profiling methods. *Front Cell Dev Biol*. 2021;9: 714687.
- Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science*. 2007;316:1497–502.
- Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*. 2007;448:553–60.
- Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*. 2007;4:651–7.
- Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. *Cell*. 2007;129:823–37.

7. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. 2013;10:1213–8.
8. Zaret K: Micrococcal nuclease analysis of chromatin structure. *Curr Protoc Mol Biol* 2005, Chapter 21:Unit 21. 21.
9. Song L, Crawford GE: DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* 2010, 2010:pdb prot5384.
10. Skene PJ, Henikoff S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife*. 2017;6: e21856.
11. Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, Ahmad K, Henikoff S. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun*. 1930;2019:10.
12. van Steensel B, Henikoff S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol*. 2000;18:424–8.
13. Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res*. 2012;22:2497–506.
14. Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science*. 2020;368:1449–54.
15. Shipony Z, Marinov GK, Swaffner MP, Sinnott-Armstrong NA, Skotheim JM, Kundaje A, Greenleaf WJ. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat Methods*. 2020;17:319–27.
16. Abdulhay NJ, McNally CP, Hsieh LJ, Kasinathan S, Keith A, Estes LS, Karimzadeh M, Underwood JG, Goodarzi H, Narlikar GJ, Ramani V. Massively multiplex single-molecule oligonucleosome footprinting. *eLife*. 2020;9: e59404.
17. Altemose N, Maslan A, Smith OK, Sundararajan K, Brown RR, Detweiler AM, Neff N, Miga KH, Straight AF, Streets A. DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome-wide. *Nat Methods*. 2022;19:711–23.
18. Liu T, Conesa A. Profiling the epigenome using long-read sequencing. *Nat Genet*. 2025;57:27–41.
19. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. 2019;37:1155–62.
20. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods*. 2010;7:461–5.
21. Douvlataniotis K, Bensberg M, Lentini A, Gylemo B, Nestor CE. No evidence for DNA N (6)-methyladenine in mammals. *Sci Adv* 2020, 6:eaay3335.
22. Bochtler M, Fernandes H. DNA adenine methylation in eukaryotes: Enzymatic mark or a form of DNA damage? *BioEssays*. 2021;43: e2000243.
23. O'Brien ZK, Boulias K, Wang J, Wang SY, O'Brien NM, Hao Z, Shibuya H, Fady PE, Shi Y, He C, et al. Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC Genomics*. 2019;20:445.
24. Kong Y, Cao L, Deikus G, Fan Y, Mead EA, Lai W, Zhang Y, Yong R, Sebra R, Wang H, et al. Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. *Science*. 2022;375:515–22.
25. Nanda AS, Wu K, Irklyenko I, Woo B, Ostrowski MS, Clugston AS, Sayles LC, Xu L, Satpathy AT, Nguyen HG, et al. Direct transposition of native DNA for sensitive multimodal single-molecule sequencing. *Nat Genet*. 2024;56:1300–9.
26. Kong Y, Mead EA, Fang G. Navigating the pitfalls of mapping DNA and RNA modifications. *Nat Rev Genet*. 2023;24:363–81.
27. Dennis AF, Xu Z, Clark DJ: Examining chromatin heterogeneity through PacBio long-read sequencing of M.EcoGII methylated genomes: an m6A detection efficiency and calling bias correcting pipeline. *Nucleic Acids Res* 2024, 52:e45.
28. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009;323:133–8.
29. Wang Y, Chen X, Sheng Y, Liu Y, Gao S. N⁶-adenine DNA methylation is associated with the linker DNA of H2A.Z-containing well-positioned nucleosomes in Pol II-transcribed genes in *Tetrahymena*. *Nucleic Acids Res*. 2017;45:11594–606.
30. Wang Y, Sheng Y, Liu Y, Zhang W, Cheng T, Duan L, Pan B, Qiao Y, Liu Y, Gao S. A distinct class of eukaryotic MT-A70 methyltransferases maintain symmetric DNA N⁶-adenine methylation at the ApT dinucleotides as an epigenetic mark associated with transcription. *Nucleic Acids Res*. 2019;47:11771–89.
31. Beh LY, Debelouchina GT, Clay DM, Thompson RE, Lindblad KA, Hutton ER, Bracht JR, Sebra RP, Muir TW, Landweber LF. Identification of a DNA N⁶-adenine methyltransferase complex and its impact on chromatin organization. *Cell*. 2019;177:1781–96.
32. Beaulaurier J, Schadt EE, Fang G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nat Rev Genet*. 2019;20:157–72.
33. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B, Mahajan MC, Jabado OJ, et al. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat Biotechnol*. 2012;30:1232–9.
34. Sheng Y, Wang Y, Yang W, Wang XQ, Lu J, Pan B, Nan B, Liu Y, Ye F, Li C, et al. Semiconservative transmission of DNA N (6)-adenine methylation in a unicellular eukaryote. *Genome Res*. 2024;34:740–56.
35. Jha A, Bohaczuk SC, Mao Y, Ranchalis J, Mallory BJ, Min AT, Hamm MO, Swanson E, Dubocanin D, Finkbeiner C, et al. DNA-m6A calling and integrated long-read epigenetic and genetic analysis with fibertools. *Genome Res* 2024;gr.279095.279124.
36. PacificBiosciences. Jasmine. GitHub. 2024. <https://github.com/PacificBiosciences/jasmine>
37. Pacific Biosciences. SMRT[®] tools reference guide (v13.0), Datasets. 2024. <https://www.pacb.com/wp-content/uploads/SMRT-Tools-Reference-Guide-v13.0.pdf>
38. Li H, Niu J, Sheng Y, Liu Y, Gao S: SMAC: identifying DNA N6-methyladenine (6mA) at the single-molecule level using SMRT CCS data. *Briefings in Bioinformatics* 2025, 26: Issue 2, bbaf153.

39. Marks P, Banerjee O, Alexander D: Detection and identification of base modifications with single molecule real-time sequencing data. GitHub 2012, <https://github.com/PacificBiosciences/kineticsTools/blob/master/doc/whitepaper/kinetics.pdf>.
40. PacificBiosciences. ipdSummary. GitHub. 2014. <https://github.com/PacificBiosciences/kineticsTools/blob/master/kineticsTools/ipdSummary.py>
41. Feng Z, Fang G, Korlach J, Clark T, Luong K, Zhang X, Wong W, Schadt E. Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. *PLoS Comput Biol*. 2013;9: e1002935.
42. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW. Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol*. 2010;472:431–55.
43. Schwartz JJ, Quake SR. Single molecule measurement of the “speed limit” of DNA polymerase. *Proc Natl Acad Sci U S A*. 2009;106:20294–9.
44. Schadt EE, Banerjee O, Fang G, Feng Z, Wong WH, Zhang X, Kislyuk A, Clark TA, Luong K, Keren-Paz A, et al. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome Res*. 2013;23:129–41.
45. PacificBiosciences. Kmer Model from ipdSummary. GitHub. 2019. <https://github.com/PacificBiosciences/kineticsTools/tree/master/kineticsTools/resources>
46. Clark TA, Murray IA, Morgan RD, Kislyuk AO, Spittle KE, Boitano M, Fomenkov A, Roberts RJ, Korlach J. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res*. 2012;40: e29.
47. Lowary PT, Widom J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J Mol Biol*. 1998;276:19–42.
48. Davey CA, Sargent DF, Luger K, Maeder AW, Richmond TJ: Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J Mol Biol* 2002, 319:1097–1113.
49. Armeev GA, Kniazeva AS, Komarova GA, Kirpichnikov MP, Shaytan AK. Histone dynamics mediate DNA unwrapping and sliding in nucleosomes. *Nat Commun*. 2021;12:2387.
50. Gansen A, Felekyan S, Kuhnemuth R, Lehmann K, Toth K, Seidel CAM, Langowski J. High precision FRET studies reveal reversible transitions in nucleosomes between microseconds and minutes. *Nat Commun*. 2018;9:4628.
51. Chua EY, Vasudevan D, Davey GE, Wu B, Davey CA. The mechanics behind DNA sequence-dependent properties of the nucleosome. *Nucleic Acids Res*. 2012;40:6338–52.
52. Vasudevan D, Chua EYD, Davey CA. Crystal structures of nucleosome core particles containing the “601” strong positioning sequence. *J Mol Biol*. 2010;403:1–10.
53. Murray IA, Morgan RD, Luyten Y, Fomenkov A, Correa IR, Jr., Dai N, Allaw MB, Zhang X, Cheng X, Roberts RJ: The non-specific adenine DNA methyltransferase M.EcoGII. *Nucleic Acids Res* 2018, 46:840–848.
54. Drozd M, Piekarczyk A, Bujnicki JM, Radlinska M. Novel non-specific DNA adenine methyltransferases. *Nucleic Acids Res*. 2012;40:2119–30.
55. Matsuo Y, MacLeod RA, Uphoff CC, Drexler HG, Nishizaki C, Katayama Y, Kimura G, Fujii N, Omoto E, Harada M, Orita K. Two acute monocytic leukemia (AML-M5a) cell lines (MOLM-13 and MOLM-14) with interclonal phenotypic heterogeneity showing MLL-AF9 fusion resulting from an occult chromosome insertion, ins(11;9)(q23;p22p23). *Leukemia*. 1997;11:1469–77.
56. Quentmeier H, Martelli MP, Dirks WG, Bolli N, Liso A, Macleod RA, Nicoletti I, Mannucci R, Pucciarini A, Bigerna B, et al. Cell line OCI/AML3 bears exon-12 NPM gene mutation-A and cytoplasmic expression of nucleophosmin. *Leukemia*. 2005;19:1760–7.
57. Baldi S, Korber P, Becker PB. Beads on a string-nucleosome array arrangements and folding of the chromatin fiber. *Nat Struct Mol Biol*. 2020;27:109–18.
58. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
59. Lai B, Gao W, Cui K, Xie W, Tang Q, Jin W, Hu G, Ni B, Zhao K. Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature*. 2018;562:281–5.
60. Hasson D, Panchenko T, Salimian KJ, Salman MU, Sekulic N, Alonso A, Warburton PE, Black BE. The octamer is the major form of CENP-A nucleosomes at human centromeres. *Nat Struct Mol Biol*. 2013;20:687–95.
61. Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ, Uralsky L, Ryabov FD, Shew CJ, et al: Complete genomic and epigenetic maps of human centromeres. *Science* 2022, 376:eabl4178.
62. Funk CC, Casella AM, Jung S, Richards MA, Rodriguez A, Shannon P, Donovan-Maiye R, Heavner B, Chard K, Xiao Y, et al. Atlas of transcription factor binding sites from ENCODE DNase hypersensitivity data across 27 tissue types. *Cell Rep*. 2020;32: 108029.
63. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. The human transcription factors. *Cell*. 2018;172:650–65.
64. Zhu F, Farnung L, Kaasinen E, Sahu B, Yin Y, Wei B, Dodonova SO, Nitta KR, Morgunova E, Taipale M, et al. The interaction landscape between transcription factors and the nucleosome. *Nature*. 2018;562:76–81.
65. Hashimoto H, Wang D, Horton JR, Zhang X, Corces VG, Cheng X. Structural basis for the versatile and methylation-dependent binding of CTCF to DNA. *Mol Cell*. 2017;66(711–720): e713.
66. Yang J, Horton JR, Liu B, Corces VG, Blumenthal RM, Zhang X, Cheng X. Structures of CTCF-DNA complexes including all 11 zinc fingers. *Nucleic Acids Res*. 2023;51:8447–62.
67. Nakahashi H, Kieffer Kwon KR, Resch W, Vian L, Dose M, Stavreva D, Hakim O, Pruett N, Nelson S, Yamane A, et al. A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep*. 2013;3:1678–89.
68. Huang H, Zhu Q, Jussila A, Han Y, Bintu B, Kern C, Conte M, Zhang Y, Bianco S, Chiariello AM, et al. CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. *Nat Genet*. 2021;53:1064–74.

69. Soochit W, Sleutels F, Stik G, Bartkuhn M, Basu S, Hernandez SC, Merzouk S, Vidal E, Boers R, Boers J, et al. CTCF chromatin residence time controls three-dimensional genome organization, gene expression and DNA methylation in pluripotent cells. *Nat Cell Biol.* 2021;23:881–93.
70. Li Y, Haarhuis JHI, Sedeno Cacciatore A, Oldenkamp R, van Ruiten MS, Willems L, Teunissen H, Muir KW, de Wit E, Rowland BD, Panne D. The structural basis for cohesin-CTCF-anchored loops. *Nature.* 2020;578:472–6.
71. Agarwal H, Reisser M, Wortmann C, Gebhardt JCM. Direct Observation of Cell-Cycle-Dependent Interactions between CTCF and Chromatin. *Biophys J.* 2017;112:2051–5.
72. Hansen AS, Pustova I, Cattoglio C, Tjian R, Darzacq X: CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *eLife* 2017, 6:e25776.
73. Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O, Okuda Y, Kawaji H, Nakaki R, Sese J, Meno C. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.* 2018;19: e46255.
74. Smith JS, Lappin KM, Craig SG, Liberante FG, Crean CM, McDade SS, Thompson A, Mills KI, Savage KI. Chronic loss of STAG2 leads to altered chromatin structure contributing to de-regulated transcription in AML. *J Transl Med.* 2020;18:339.
75. Michel BC, D'Avino AR, Cassel SH, Mashtalir N, McKenzie ZM, McBride MJ, Valencia AM, Zhou Q, Bocker M, Soares LMM, et al. A non-canonical SWI/SNF complex is a synthetic lethal target in cancers driven by BAF complex perturbation. *Nat Cell Biol.* 2018;20:1410–20.
76. Antony C, George SS, Blum J, Somers P, Thorsheim CL, Wu-Corts DJ, Ai Y, Gao L, Lv K, Tremblay MG, et al. Control of ribosomal RNA synthesis by hematopoietic transcription factors. *Mol Cell.* 2022;82(3826–3839): e3829.
77. Heberle E, Bardet AF. Sensitivity of transcription factors to DNA methylation. *Essays Biochem.* 2019;63:727–41.
78. Frederick MA, Williamson KE, Fernandez Garcia M, Ferretti MB, McCarthy RL, Donahue G, Luzete Monteiro E, Takenaka N, Reynaga J, Kadoch C, Zaret KS. A pioneer factor locally opens compacted chromatin to enable targeted ATP-dependent nucleosome remodeling. *Nat Struct Mol Biol.* 2023;30:31–7.
79. De la Rica L, Rodriguez-Ubrea J, Garcia M, Islam AB, Urquiza JM, Hernando H, Christensen J, Helin K, Gomez-Vaquero C, Ballestar E: PU.1 target genes undergo Tet2-coupled demethylation and DNMT3b-mediated methylation in monocyte-to-osteoclast differentiation. *Genome Biol* 2013, 14:R99.
80. Joshi K, Liu S, Breslin SJP, Zhang J. Mechanisms that regulate the activities of TET proteins. *Cell Mol Life Sci.* 2022;79:363.
81. Kim S, Wysocka J. Deciphering the multi-scale, quantitative cis-regulatory code. *Mol Cell.* 2023;83:373–92.
82. Tullius TW, Isaac RS, Dubocanin D, Ranchalis J, Churchman LS, Stergachis AB. RNA polymerases reshape chromatin architecture and couple transcription on individual fibers. *Mol Cell.* 2024;84(3209–3222): e3205.
83. Falkenberg M, Larsson NG, Gustafsson CM. Replication and Transcription of Human Mitochondrial DNA. *Annu Rev Biochem.* 2024;93:47–77.
84. Kopinski PK, Singh LN, Zhang S, Lott MT, Wallace DC. Mitochondrial DNA variation and cancer. *Nat Rev Cancer.* 2021;21:431–45.
85. Wei W, Schon KR, Elgar G, Orioli A, Tanguy M, Giess A, Tischkowitz M, Caulfield MJ, Chinnery PF. Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. *Nature.* 2022;611:105–14.
86. Isaac RS, Tullius TW, Hansen KG, Dubocanin D, Couvillion M, Stergachis AB, Churchman LS. Single-nucleoid architecture reveals heterogeneous packaging of mitochondrial DNA. *Nat Struct Mol Biol.* 2024;31:568–77.
87. Brown TA, Cecconi C, Tkachuk AN, Bustamante C, Clayton DA. Replication of mitochondrial DNA occurs by strand displacement with alternative light-strand origins, not via a strand-coupled mechanism. *Genes Dev.* 2005;19:2466–76.
88. Peter B, Falkenberg M. TWINKLE and other human mitochondrial DNA helicases: structure, function and disease. *Genes (Basel).* 2020;11:408.
89. Milenkovic D, Matic S, Kuhl I, Ruzzenente B, Freyer C, Jemt E, Park CB, Falkenberg M, Larsson NG. TWINKLE is an essential mitochondrial helicase required for synthesis of nascent D-loop strands and complete mtDNA replication. *Hum Mol Genet.* 2013;22:1983–93.
90. Lee YS, Kennedy WD, Yin YW. Structural insight into processive human mitochondrial DNA synthesis and disease-related polymerase mutations. *Cell.* 2009;139:312–24.
91. Zhang J, Peng Q, Ma C, Wang J, Xiao C, Li T, Liu X, Zhou L, Xu X, Zhou WZ, et al: 6mA-Sniper: Quantifying 6mA sites in eukaryotes at single-nucleotide resolution. *Sci Adv* 2023, 9:eadh7912.
92. Berman AJ, Kamtekar S, Goodman JL, Lazaro JM, de Vega M, Blanco L, Salas M, Steitz TA. Structures of phi29 DNA polymerase complexed with substrate: the mechanism of translocation in B-family polymerases. *EMBO J.* 2007;26:3494–505.
93. 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. *Nat Methods.* 2020;17:1191–9.
94. Wang Y, Wang A, Liu Z, Thurman AL, Powers LS, Zou M, Zhao Y, Hefel A, Li Y, Zabner J, Au KF. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Res.* 2019;29:1329–42.
95. Lai WKM, Pugh BF. Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nat Rev Mol Cell Biol.* 2017;18:548–62.
96. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* 2005;33:5868–77.
97. Hung KL, Luebeck J, Dehkordi SR, Colon CI, Li R, Wong IT, Coruh C, Dharanipragada P, Lomeli SH, Weiser NE, et al. Targeted profiling of human extrachromosomal DNA by CRISPR-CATCH. *Nat Genet.* 2022;54:1746–54.
98. Gilman P, Janzou S, Guittet D, Freeman J, DiOrto N, Blair N, Boyd M, Neises T, Wagner M: PySAM (Python wrapper for system advisor model "SAM") [SWR-19–57]. Computer software 2019, <https://github.com/NREL/pysam>.
99. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics.* 2012;13:1–18.
100. Yang, W, Liu, Y, Dou, Y. ipdTrimming: 6mA calling pipeline. GitHub. 2024. <https://github.com/YifanLab/ipdTrimming>
101. Portik D, Hall R, Nyquist K, Wenger A: Extracting CpG methylation from PacBio HiFi whole genome sequencing. Pacific Biosciences, Menlo Park, CA 2022, <https://www.pacb.com/wp-content/uploads/AGBT-2022-extracting-CpG-methylation-Portik-poster.pdf>.

102. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. *SciPy*. 2010;57:10–25080.
103. Moritz S, Bartz-Beielstein T. imputeTS: time series missing value imputation in R. *R J*. 2017;9:207.
104. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J: SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* 2020, 17:261–272.
105. Nurk S, Koren S, Rhie A, Rautiainen M, Bizikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A. The complete sequence of a human genome. *Science*. 2022;376:44–53.
106. Ye J, McGinnis S, Madden TL. BLAST: improvements for better sequence analysis. *Nucleic Acids Res*. 2006;34:W6–9.
107. Suzuki Y, Myers EW, Morishita S: Rapid and ongoing evolution of repetitive sequence structures in human centromeres. *Sci Adv* 2020, 6:eabd9230.
108. Suzuki, Y. Codes for centromere analysis. GitHub. 2020. https://github.com/hacone/hc_temporary/releases/tag/submission-v.1.0
109. Fornes O, Castro-Mondragon JA, Khan A, Van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2020;48:D87–92.
110. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. 2011;27:1017–8.
111. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2.
112. Chris Benner. HOMER v4.11.1. Datasets. 2019. <http://homer.ucsd.edu/homer/data/genomes/hg19.v6.4.zip>
113. Aurélien Guy-Duché, Pierre Lindenbaum. Blast2Bam. GitHub. 2021. <https://github.com/guyduche/Blast2Bam>.
114. Sha L, Yang Z, An S, Yang W, Kim S, Oh H, Xu J, Yin J, Wang H, Lenz HJ, et al. Non-canonical MLL1 activity regulates centromeric phase separation and genome stability. *Nat Cell Biol*. 2023;25:1637–49.
115. Yang, W, Liu, Y, Dou, Y. Revealing long-range heterogeneous organization of nucleoproteins with N6-methyladenine footprinting by ipdTrimming. Datasets. 2025. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1166413>
116. Yang W, Liu Y, Dou Y. Revealing long-range heterogeneous organization of nucleoproteins with N6-methyladenine footprinting by ipdTrimming. 2025. Zenodo. <https://doi.org/10.5281/zenodo.14602961>.
117. Sheng Y. Semiconservative transmission of DNA N (6)-adenine methylation in a unicellular eukaryote. Datasets. 2024. <https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos3/sra-pub-run-36/SRR016/16944/SRR16944761/SRR16944761.1>
118. Sheng Y. Semiconservative transmission of DNA N (6)-adenine methylation in a unicellular eukaryote. Datasets. 2024. <https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos3/sra-pub-run-36/SRR028/28602/SRR28602303/SRR28602303.1>
119. Sheng Y. Semiconservative transmission of DNA N (6)-adenine methylation in a unicellular eukaryote. Datasets. 2024. <https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos3/sra-pub-run-36/SRR017/17010/SRR17010848/SRR17010848.1>
120. Abdulhay, N. Massively multiplex single-molecule oligonucleosome footprinting. Datasets. 2020. <https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos3/sra-pub-zq-24/SRR013/13173/SRR13173521/SRR13173521.lite.1>
121. Sheng Y. Semiconservative transmission of DNA N (6)-adenine methylation in a unicellular eukaryote. Datasets. 2024. <https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos3/sra-pub-run-36/SRR027/27430/SRR27430828/SRR27430828.1>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.