Briefings in Bioinformatics, 2025, **26(2)**, bbaf153
https://doi.org/10.1093/bib/bbaf153
Problem Solving Protocol

SMAC: identifying DNA N^6-methyladenine (6mA) at the single-molecule level using SMRT CCS data

Haicheng Li[1,2,‡], Junhua Niu[1,2,‡], Yalan Sheng[3,‡], Yifan Liu[4,*], Shan Gao[1,2,*]

[1]MOE Key Laboratory of Evolution & Marine Biodiversity and Institute of Evolution & Marine Biodiversity, Ocean University of China, 5 Yushan Road, Qingdao 266003, China
[2]Laboratory for Marine Biology and Biotechnology, Qingdao Marine Science and Technology Center, 168 Wenhai Middle Road, Qingdao 266237, China
[3]Shum Yiu Foon Shum Bik Chuen Memorial Centre for Cancer and Inflammation Research, School of Chinese Medicine, Hong Kong Baptist University, 7 Baptist University Road, Kowloon Tong, Hong Kong 999077, China
[4]Department of Biochemistry & Molecular Medicine, University of Southern California Keck School of Medicine, 1441 Eastlake Avenue, Los Angeles, CA 90033, United States
*Corresponding authors. Yifan Liu, Department of Biochemistry & Molecular Medicine, University of Southern California Keck School of Medicine, 1441 Eastlake Avenue, Los Angeles, CA 90033, USA. E-mail: Yifan.Liu@med.usc.edu; Shan Gao, MOE Key Laboratory of Evolution & Marine Biodiversity and Institute of Evolution & Marine Biodiversity, Ocean University of China, 5 Yushan Road, Qingdao 266003, China. E-mail: shangao@ouc.edu.cn
‡Haicheng Li, Junhua Niu and Yalan Sheng contributed equally.

Abstract

DNA modifications, such as N^6-methyladenine (6mA), play important roles in various processes in eukaryotes. Single-molecule, real-time (SMRT) sequencing enables the direct detection of DNA modifications without requiring special sample preparation. However, most SMRT-based studies of 6mA rely on ensemble-level consensus by combining multiple reads covering the same genomic position, which misses the single-molecule heterogeneity. While recent methods have aimed at single-molecule level detection of 6mA, limitations in sequencing platforms, resolution, accuracy, and usability restrict their application in comprehensive epigenetic studies. Here, we present SMAC (single-molecule 6mA analysis of CCS reads), a novel framework for accurately detecting 6mA at the single-molecule level using SMRT circular consensus sequencing (CCS) data from the Sequel II system. It is an automated method that streamlines the entire workflow by packaging both existing softwares and built-in scripts, with user-defined parameters to allow easy adaptation for various studies. By utilizing the statistical distribution characteristics of enzyme kinetic indicators on single DNA molecules rather than a fixed cutoff, SMAC significantly improves 6mA detection accuracy at the single-nucleotide and single-molecule levels. It simplifies analysis by providing comprehensive information, including quality control, statistical analysis, and site visualization, directly from raw sequencing data. SMAC is a powerful new tool that enables *de novo* detection of 6mA and empowers investigation of its functions in modulating physiological processes.

Keywords: DNA N^6-methyladenine (6mA); single molecule; SMRT CCS sequencing; SMAC (single-molecule 6mA analysis of CCS reads)

Introduction

DNA modifications, such as N^6-methyladenine (6mA), play significant roles in various crucial processes in eukaryotes, including DNA structure maintenance, transcriptional regulation, nucleosome positioning, transposon activation, and DNA replication. 6mA is closely associated with stress responses, embryonic development, cellular physiological states, tumor cell growth, plant growth and development, as well as immune responses [1–12]. Understanding the distribution of 6mA in eukaryotic genomes is essential for uncovering its regulation mechanisms and biological implications.

Several methodologies have been developed to detect 6mA, including dot blot, liquid chromatography–mass spectrometry (LC–MS), and 6mA immunoprecipitation sequencing (6mA-IP-seq) [13–15]. While these techniques have provided valuable insights, each comes with specific limitations. For instance, dot blot and LC–MS can estimate total 6mA levels, but they provide no information of context sequences and cannot rule out contamination from DNA of other species [13, 15]. Additionally, 6mA-IP-seq, which relies on Illumina short-fragment sequencing, can identify 6mA-enriched genomic regions but lacks single-nucleotide resolution [13–15]. Moreover, both dot blot and 6mA-IP-seq may produce false-positive signals due to non-specific antibody binding [13, 15]. In contrast, single-molecule, real-time (SMRT) sequencing offers the direct detection of DNA modifications at the single-base resolution across long genomic regions [13–15], making it a powerful tool for detecting 6mA with unprecedented accuracy and detail.

During SMRT sequencing, double-stranded native DNA fragments are circularized by ligating hairpin adapters to each end. DNA polymerase then proceeds around the circularized DNA multiple times, with the number of passes depending on the fragment size and polymerase processivity [16, 17]. The time taken by the polymerase to translocate from one nucleotide to the next is termed the inter-pulse duration (IPD), and variations in IPDs are highly correlated with modifications in the DNA template [17–19]. The modification is detected by calculating the IPD ratio between the IPD values of tested samples and those of a whole genome

Received: November 26, 2024. **Revised:** February 17, 2025. **Accepted:** March 16, 2025

amplification (WGA) control or an *in silico* negative control provided by the sequencing platform. SMRT sequencing can be performed in two modes: continuous long read (CLR) and circular consensus sequencing (CCS). While CLR is effective for mapping 6mA at the ensemble level by combining data from different DNA molecules covering the same genomic position, it lacks the ability to provide single-molecule information. In contrast, CCS, with shorter fragment sizes and improved polymerase processivity, enables more passes (subreads) over the same DNA molecule, thus generating high-fidelity (HiFi) reads with improved sequence accuracy [20–22]. More importantly, subreads are combined to accurately call 6mA at single-molecule levels, allowing for exploration of heterogeneities among different molecules [23–25].

The classic standard SMRT-seq pipeline, such as the ipdSummary tool in the SMRT Link software, can be applied to both CLR and CCS but has typically been used to detect 6mA at the ensemble level [17, 26]. While frameworks like single-molecule modification analysis of long reads (SMALR) have been developed for single-molecule 6mA detection, they were designed for earlier CCS data produced by the PacBio RSII system [23]. In addition, its case–control method requires sequencing methylation-free samples, significantly increasing costs. With advancements in PacBio SMRT CCS technology, the datasets generated by the Sequel II system have increased tenfold in size. More recent CCS-based methods, such as 6mASCOPE, offer higher sensitivity but do not output results at the single-nucleotide resolution, instead providing only an assessment of the overall 6mA/A level [27]. Another approach, 6mA-Sniper, enables single-molecule 6mA detection but relies on fragmented scripts that are challenging to operate and provides no data analysis and visualization [28]. Due to limitations in current 6mA detection tools and pipelines base on CCS data, many studies still rely on CLR data, or CCS data analyzed using the standard SMRT-seq pipeline [7, 11, 12, 29–33], failing to capture 6mA methylation states of individual molecules.

In this study, we present SMAC (single-molecule 6mA analysis of CCS reads), a novel and automated toolkit designed for single-molecule, single-nucleotide, and strand-specific detection of 6mA using SMRT CCS data. SMAC is built on the continuously updated PacBio SMRT Link software and is compatible with the commonly used SMRT sequencing platform. Unlike SMALR that requires additional methylation-free datasets, SMAC employs *in silico* controls embedded in ipdSummary. It uses molecule-specific IPD ratio information to infer methylation states at the single-molecule resolution. SMAC offers several key advantages: First, it applies rigorous data pretreatment to minimize background noise. Second, it identifies more reliable genome-wide 6mA sites than the standard SMRT-seq pipeline while also detecting 6mA at the single-molecule level. Third, it enhances the accuracy of 6mA detection within ApT motifs, where methylation can occur full- or hemi-methylated. Fourth, by using a Gaussian distribution fitting approach, SMAC allows for a more objective determination of the cutoff for 6mA site detection. Last, SMAC requires no bioinformatics expertise; users can receive quality control information, statistical analysis, and visualized results simply by inputting raw sequencing data. SMAC is freely available on GitHub (https://github.com/liiihc/SMAC).

Materials and methods

SMAC includes multiple steps such as data preprocessing, alignment, and 6mA identification. The raw subreads data are processed using the "ccs" module in SMRT Link to generate HiFi reads with the parameter "–hifi-kinetics" and converted to FASTA format using the "bam2fasta" module, which also provides a quality report for the sequencing run. By default, only reads with ≥20× passes are retained for downstream analysis. Users can modify the minimal number of passes by adjusting the "-passes" parameter within the SMAC toolkit, depending on their need for either larger datasets or stricter quality control. The HiFi reads are split into individual FASTA files to serve as reference sequences. The raw subreads are converted to SAM format using SAMtools and further split into individual SAM files for analysis. A reference index of the split HiFi reads is built using the "pbindex" module in SMRT Link. Each SAM file is converted to BAM format and aligned to the corresponding HiFi reads using the "pbmm2" module with the default parameter, while the IPD ratio is calculated using the "ipdSummary" module.

The HiFi reads are aligned to the reference genome using both BLASTN and "pbmm2" with only reads meeting the criteria of ≥80% coverage and ≥80% identity in the BLASTN results being retained for further analysis by default. Users can customize the BLASTN coverage and identity thresholds by modifying the "-coverage" and "-identity" parameters. To ensure accuracy, the IPD ratios of bases within 25 bp of the adapter sequences are trimmed by default. Users can modify the number of trimmed bases based on the IPD ratio variation in their own dataset, by adjusting the "-trimmer" parameter. Each adenine (A) in the CCS reads is mapped back to the genome based on the "pbmm2" alignment results. The IPD ratio distribution of all adenines aligned to the reference genome is calculated, and a Gaussian distribution is fitted to determine the initial cutoff. By default, only reads with standard deviation of IPD ratios ≤0.6 for non-6mA bases on both Watson and Crick strands are retained for downstream analysis. Users can modify the cutoff based on the IPD ratio variation in their own dataset, by adjusting the "-sd_cutoff" parameter. For ApT dinucleotides, a second Gaussian fitting is applied to the IPD ratios of the corresponding adenine within the initially identified 6mApT sites to determine a secondary cutoff. Using these cutoffs, 6mA sites are distinguished from non-methylated adenines at the single-molecule level. Users can modify the "-ipd_cutoff" parameter to apply cutoff determined by the Gaussian fit or any specific cutoff. The penetrance of each 6mA site is defined as the ratio of the number of 6mA sites to the total number of adenines across all reads.

Users are required to have BLASTN, SMRT Link, perl module "Statistics::Descriptive," and python packages, including "logomaker lmfit, scipy, statsmodels, pandas, matplotlib, and numpy," installed in advance.

Results and discussion

Conceptual view of single-molecule 6mA analysis of CCS reads

Figure 1 outlines the conceptual framework of SMAC, which uses the IPD ratio from the CCS mode of the Sequel II system to call 6mA at the single-molecule level. The process starts with generating HiFi reads from raw subreads data using the ccs module in SMRT Link (Pacific Biosciences). Subreads and HiFi reads are then split based on single-molecule IDs into single molecules, filtering out those with low passes. Next, subreads are aligned to their corresponding HiFi reads using the pbmm2 module in SMRT Link. The IPD ratios for all adenine sites are computed using the ipdSummary module in SMRT Link. Before calling 6mA, HiFi reads are mapped to the reference genome using BLASTN and pbmm2 to filter out reads originating from contaminants and locate the bases within the genome. A Gaussian distribution is then fitted to

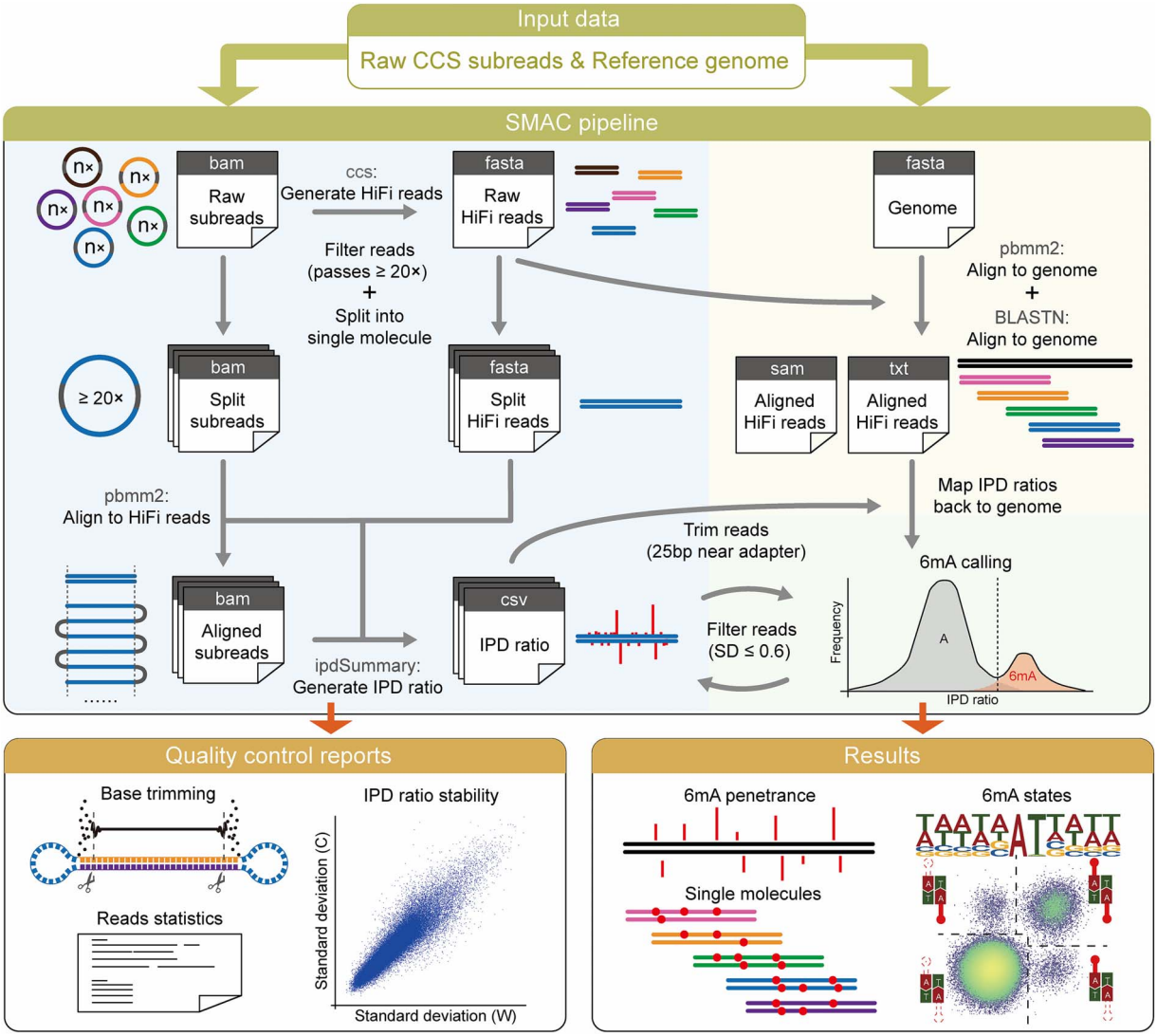

Figure 1. Conceptual view of SMAC. Overview of SMAC for detecting 6mA from SMRT CCS data. SMAC is an automated toolkit that processes original subreads files and the reference genome file of the target species, generating quality control reports and results of both ensemble and single-molecule level 6mA sites, along with the methylation states for each individual ApT dinucleotide. CCS, circular consensus sequencing. IPD, inter-pulse duration. Penetrance, the proportion of molecules containing 6mA at a specific genomic position.

separate the 6mA peak from the background, establishing a cutoff value of IPD ratio to identify 6mA sites and assess the methylation states of 6mA at A on a single-molecule basis. During the data processing, SMAC provides quality control reports for base trimming, IPD ratio stability, and basic statistics for library size and polymerase processivity. The final output includes 6mA calling results at both ensemble and single-molecule levels, along with the methylation states for each individual ApT dinucleotide. Most analyses presented in this study were based on native genomic DNA sample from the wild-type (WT, SB210 strain) *Tetrahymena thermophila*, an important unicellular model eukaryote with abundant 6mA in its genome [4, 6, 7, 10, 34, 35], unless otherwise noted.

SMAC packages all required steps into a comprehensive and flexible toolkit, enabling users to obtain analysis results simply by providing the original subreads files and the reference genome file of the target species. It also allows users to adjust key parameters to suit their specific needs. Additionally, SMAC generates result files compatible with the Integrative Genomics Viewer (https://igv.org/) (Supplementary Fig. S1), enabling users to visualize both ensemble and single-molecule 6mA information.

Data trimming and filtration

SMRT CCS sequencing involves circular sequencing of SMRTbell templates, which consist of DNA molecules with adapters attached at both ends. To investigate whether the presence of these adapters can influence the IPD ratios near the ends of DNA molecules, we analyzed a published CCS dataset of WGA for WT *T. thermophila* [10]. By examining the IPD ratio distribution within a defined distance from the molecule ends, we observed increased variability in the bases within 25 bp of adapters (Fig. 2a). Additionally, abnormal bases with an IPD ratio of zero were frequently observed in this region for unknown reasons (Supplementary Fig. S2A). Based on these findings, SMAC recommends trimming 25 bp from both ends of DNA molecules to minimize interference as a default setting.

The number of passes, or the number of subreads for each DNA molecule, significantly impacts the stability of IPD detection, with higher pass numbers resulting in more reliable IPD ratios. We categorized all molecules based on their number of passes and calculated the standard deviation of IPD ratios for bases on both the Watson and Crick strands, for the WGA sample. As the number

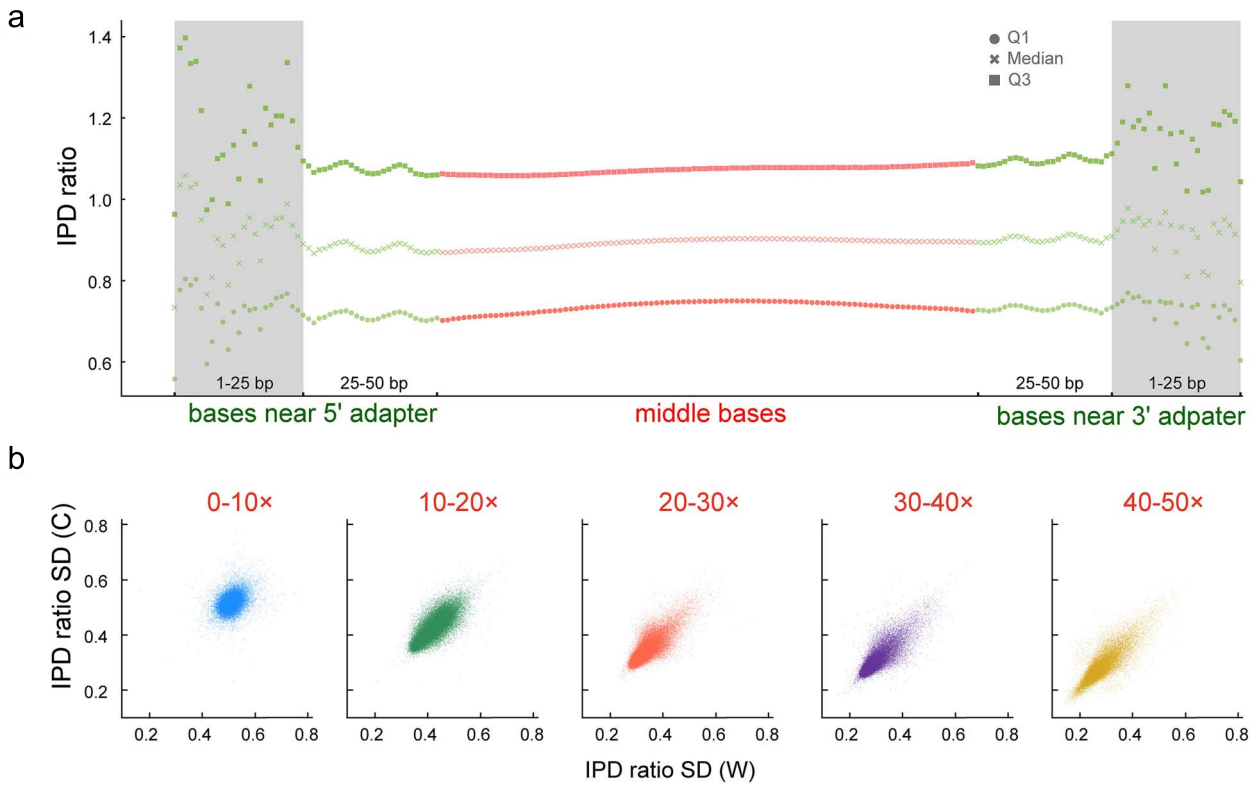

Figure 2. Data trimming and filtration. (a) Distribution of IPD ratios in SMRT CCS reads. IPD ratios on bases within 50 bp near the 5′ or 3′ adapter were shown as green dots. Other bases located in the central region were normalized to a scale of 100 units and represented by red dots. Q1, median, and Q3 represent the first quartile, median, and the third quartile of IPD ratios. (b) Average standard deviation (SD) of IPD ratios for SMRT CCS reads across different pass ranges: 0–10×, 10–20×, 20–30×, 30–40×, and 40–50×. SD values were calculated separately for IPD ratios on the Watson strand (W) and Crick strand (C).

of passes increased, the standard deviation decreased, stabilizing at about 20 passes (Fig. 2b). Therefore, SMAC recommends filtering out molecules with fewer than 20 passes as a default setting.

Contamination from external DNA poses a significant challenge for 6mA detection, including mass spectrometry and dot blot assays. Even in SMRT sequencing, a few chimeric DNA molecules formed during library preparation can introduce considerable contamination, especially from bacterial DNA with high levels of 6mA. To address this, SMAC employs BLASTN to filter out chimeric molecules. We used a mixed sample of *Saccharomyces cerevisiae* (target species), *Escherichia coli* (contamination), and *Helicobacter pylori* (contamination) to examine the effect of BLASTN on reads mapping ratio and chimeric reads ratio. As the requirements for coverage and identity in BLASTN increased, the chimeric reads ratio gradually decreased, but the ratio of mapped reads retained after BLASTN alignment also significantly declined (Supplementary Fig. S2B). To minimize the proportion of chimeric reads while retaining as much valid data as possible, we established default BLASTN filtering parameters of 80 for both coverage and identity.

Sensitive and accurate 6mA calling

In SMRT sequencing, 6mA exhibits higher IPD ratios compared to unmethylated adenines (A), leading to a bimodal distribution of IPD ratios in native DNA sample. In contrast, WGA sample lacking methylation displays a single peak corresponding to A (Fig. 3a). SMAC fits a Gaussian distribution to the high IPD ratio peak for native DNA sample, with the intersection of this fitted curve and the residual curve serving as the cutoff for 6mA detection (Fig. 3a). Sites with IPD ratios exceeding this cutoff are identified as 6mA.

SMAC aligns sequences of individual molecules back to the genome and calculates the penetrance, which represents the proportion of molecules containing 6mA at a specific genomic position (Fig. 3b). In native DNA sample, SMAC detected 652 831 shared 6mA sites with the standard SMRT-seq pipeline based on SMRT Link, accounting for 81.02% of total 6mA sites identified by SMAC (Fig. 3c). Additionally, SMAC identified 152 295 unique sites (18.98%) that the standard SMRT-seq pipeline failed to detect (Fig. 3c). Conversely, the standard SMRT-seq pipeline detected only 8324 unique sites (1.03% of total 6mA sites identified by both SMAC and the standard SMRT-seq pipeline) (Fig. 3c), demonstrating an overall higher 6mA/A level calculated by SMAC (SMAC: 0.74%, the standard SMRT-seq pipeline: 0.63%). For the WGA sample lacking a bimodal distribution, SMAC identified no valid 6mA sites, as there was no 6mA peak for Gaussian fitting. In contrast, the standard SMRT-seq pipeline falsely identified 9028 sites as 6mA (Supplementary Fig. S3A). Together, SMAC identified more high-confidence 6mA sites while maintaining a lower false-positive rate.

Regarding the assessment of 6mA levels, the penetrance calculated by SMAC correlated well with the methylation level determined by the standard SMRT-seq pipeline for shared 6mA sites. Although SMAC-specific sites displayed relatively lower penetrance (Fig. 3d and Supplementary Fig. S3B), they demonstrated high coverage, reflecting reliable sequencing depth (Supplementary Fig. S3B). More importantly, these SMAC-specific sites exhibited typical features of authentic 6mA, as

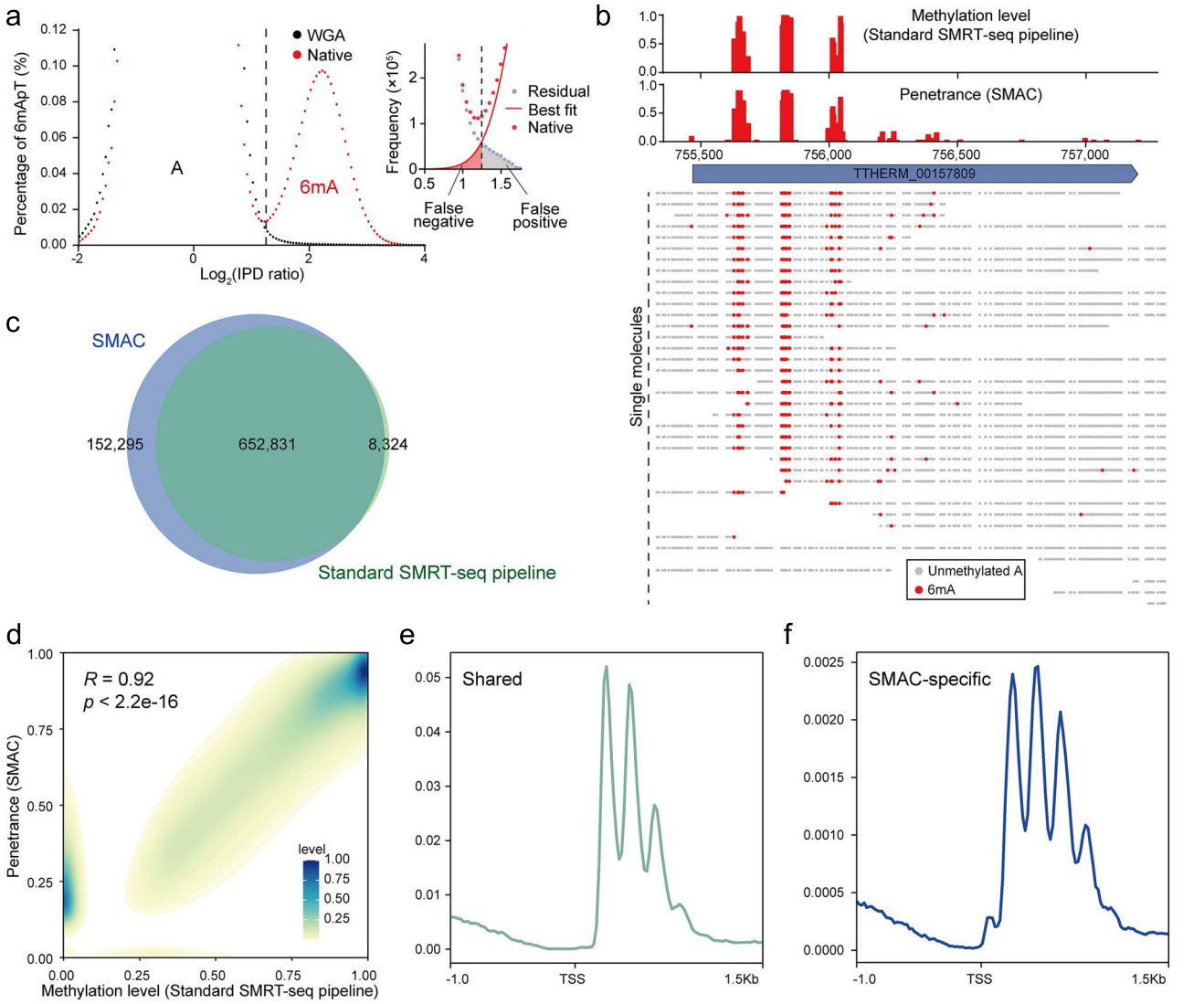

Figure 3. Sensitive and accurate 6mA calling. (a) IPD ratio distribution (Log_2) of A sites in native DNA and WGA sample. Note the deconvolution of the 6mA peak (right) and the unmodified A peak (left) in IPD ratio distribution. Residual represented the difference between authentic IPD distribution and the best fit curve of Gaussian distribution in the small 6mA peak. The IPD ratio cutoff was indicated by a dashed line. (b) Comparison of the 6mA methylation level calculated by the standard SMRT-seq pipeline and penetrance calculated by SMAC, showcased by a representative genomic region from the native DNA sample. 6mA and unmodified A sites on each individual DNA molecules were also displayed. (c) Overlap between 6mA sites called by SMAC (6mA coverage ≥5×) and the standard SMRT-seq pipeline (coverage ≥100, Qv ≥30). (d) Correlation between 6mA methylation level calculated by the standard SMRT-seq pipeline and penetrance calculated by SMAC. 6mA site distribution density was plotted as a heat map. (e) Distribution profiles of 6mA sites identified by both the standard SMRT-seq pipeline and SMAC. 6mA peaks had a periodicity distribution downstream of TSS. TSS, transcription start site. (f) Distribution profiles of SMAC-specific 6mA sites.

demonstrated by shared 6mA sites. First, they had a strong preference for the ApT dinucleotide (Supplementary Table S1). Second, they were predominantly located downstream of transcription start sites (TSS) and displayed a periodic distribution between nucleosome arrays (Fig. 3e and f and Supplementary Fig. S3D) [4, 6, 7]. As a control, those false-positive sites in the WGA sample identified by the standard SMRT-seq pipeline also had high coverage but lacked these features (Supplementary Fig. S3C and D).

Based on these comparisons, we conclude that SMAC can identify authentic 6mA with higher sensitivity and accuracy than the standard SMRT-seq pipeline.

Determination of 6mApT states

In most prokaryotes and unicellular eukaryotes, 6mA typically occurs at the self-complementary ApT dinucleotide, existing in either a full- or hemi-methylated state [1, 4, 12, 18, 27, 36, 37]. These states may have distinct downstream regulatory functions, highlighting the importance of accurately distinguishing them [11, 12, 32]. SMAC performs an initial Gaussian fitting of IPD ratios for all A bases to generate the first cutoff for calling methylated ApT (6mApT) (Fig. 4a). Moreover, since the IPD ratio distribution pattern of A on the opposite strand of 6mApT identified after the initial Gaussian fitting slightly changed, the IPD ratio distribution of the opposing A within initially identified 6mApT sites is recalculated. A second Gaussian fitting is then conducted to generate the second cutoff to refine the 6mApT calling (Fig. 4a). After two rounds of 6mA calling, ApT are classified into four groups: full-methylated, hemi-methylated on the Watson strand (hemi-W), hemi-methylated on the Crick strand (hemi-C), and unmethylated (Fig. 4b).

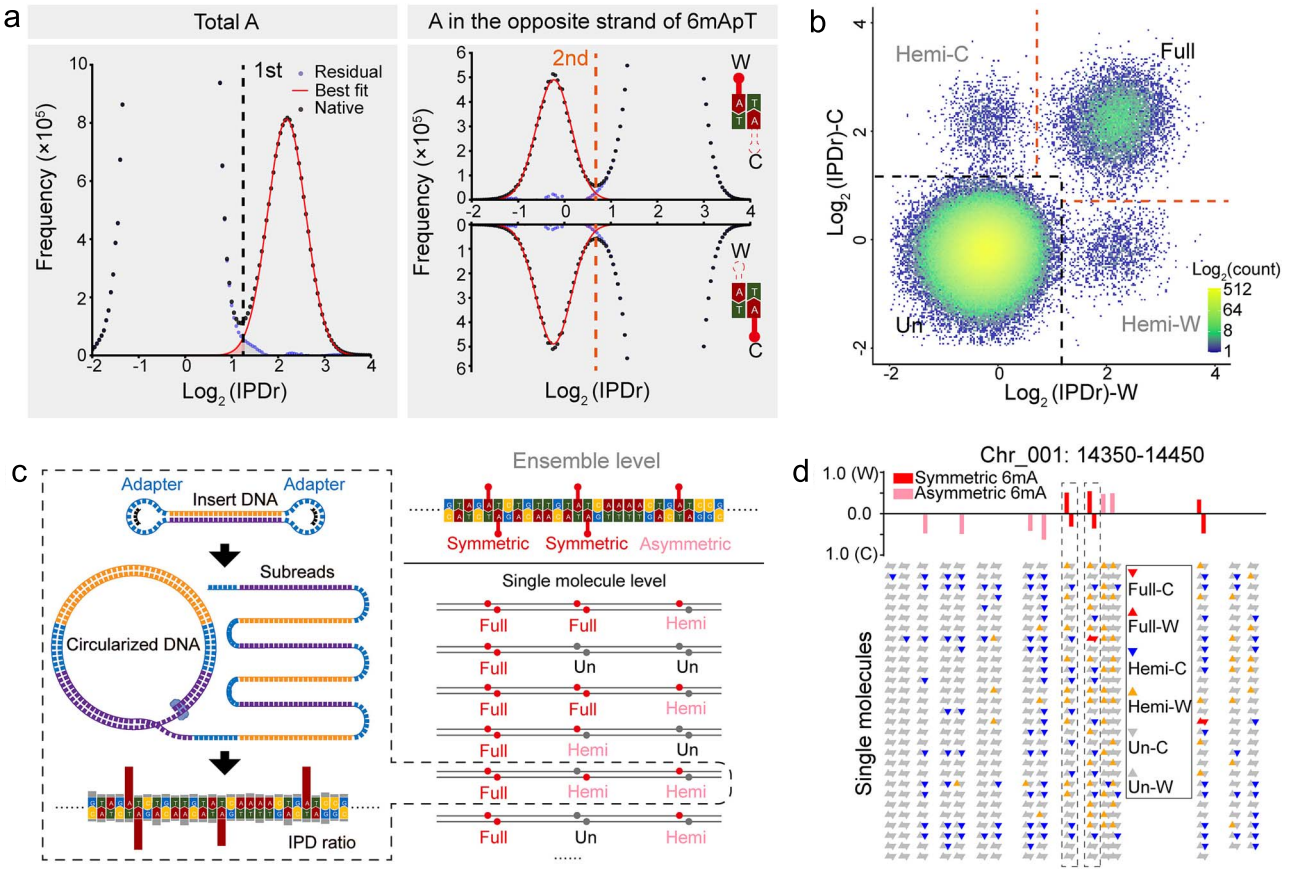

Figure 4. Determination of 6mApT states. (a) Deconvolution of the 6mA peak and the unmodified A peak for IPD ratio distribution. The left panel showed the initial Gaussian fitting of the small 6mA peak of all A. The right panel showed the second Gaussian fitting result of the unmodified A of the opposing A within initially identified 6mApT. Residual was the difference between authentic IPD distribution and the best fit curve. The IPD ratio cutoff was marked by dashed line. (b) Demarcation of the four methylation states of ApT duplexes by their IPD ratio on Watson (W) and Crick (C) strands, respectively. 6mA site distribution density was plotted as a heat map. The IPD ratio cutoff for 6mA calling was set according to deconvolution based on two rounds of Gaussian fitting. The IPD ratio cutoff was marked by dashed line. (c) Schematic diagram illustrating the determination of 6mApT states at ensemble and single-molecule levels. The left panel showed the 6mA detection of a CCS single DNA molecule. The right panel showed the discrepancy between full/hemi sites at the single-molecule level and symmetric/asymmetric methylation at the ensemble level. (d) Representative DNA molecules from the maintenance methyltransferase knockout *Tetrahymena* strain Δ*AMT1*. The 6mA states as symmetric or asymmetric at the ensemble level identified by the standard SMRT-seq pipeline (top) were largely different from the full or hemi states revealed by SMAC (bottom). ApT dinucleotides with distinct methylation states interspersed with unmodified ApTs.

Full and hemi-methylation at the single-molecule level correspond to symmetric and asymmetric methylation at the ensemble level, respectively. However, the standard SMRT-seq pipeline tends to overestimate symmetric methylation by misinterpreting multiple hemi-methylated molecules as symmetric, whereas SMAC, by leveraging single-molecule information, avoids this issue (Fig. 4c). We used both SMAC and the standard SMRT-seq pipeline to compare two *T. thermophila* samples: one predominantly featuring full methylation (WT, 89.29% full-6mApT) and another primarily exhibiting hemi-methylation (maintenance methyltransferase knockout strain Δ*AMT1*, 97.63% hemi-6mApT) (Supplementary Table S2) [10]. In both samples, the standard SMRT-seq pipeline revealed a much higher proportion of symmetric methylation compared to the full methylation level identified by SMAC, a discrepancy particularly pronounced in the sample dominated by hemi-methylation (Fig. 4d and Supplementary Table S2). Overall, SMAC outperforms the standard SMRT-seq pipeline in accurately determining 6mApT states.

Detection limit and minimal sequencing depth

SMAC determines the threshold for 6mA detection by fitting a Gaussian curve to the bimodal distribution of IPD ratios. This strategy may be too stringent for samples with low methylation levels, as they might not display a clear bimodal pattern. To explore this issue, we diluted a native DNA sample (6mA/A = 0.74%) with a WGA sample to various levels and applied the bimodal model for curve fitting. The error rate for calling 6mA increased significantly when the 6mA/A level approached approximately 0.1%. When the 6mA/A level dropped below 0.04%, the IPD ratios of A no longer exhibited a bimodal distribution, with the 6mA signal on the right side completely masked by the noise from A on the left side (Fig. 5a). Therefore, SMAC can reliably detect 6mA in samples where 6mA/A levels exceed 0.04%.

For organisms with low 6mA levels, such as most multicellular eukaryotes (e.g., *Homo sapiens*, *Arabidopsis thaliana*, *Drosophila melanogaster*, and *Caenorhabditis elegans*), SMAC is unable to identify 6mA due to the absence of a bimodal distribution in their IPD ratios (Supplementary Fig. S4). In contrast, SMAC can successfully detect 6mA in prokaryotes and lower unicellular eukaryotes, such as *E. coli*, *Chlamydomonas reinhardtii* (Supplementary Fig. S4) and *T. thermophila*. Notably, for organisms with high 6mApT abundance, SMAC offers a more precise determination of 6mApT states (Fig. 4 and Supplementary Table S2).

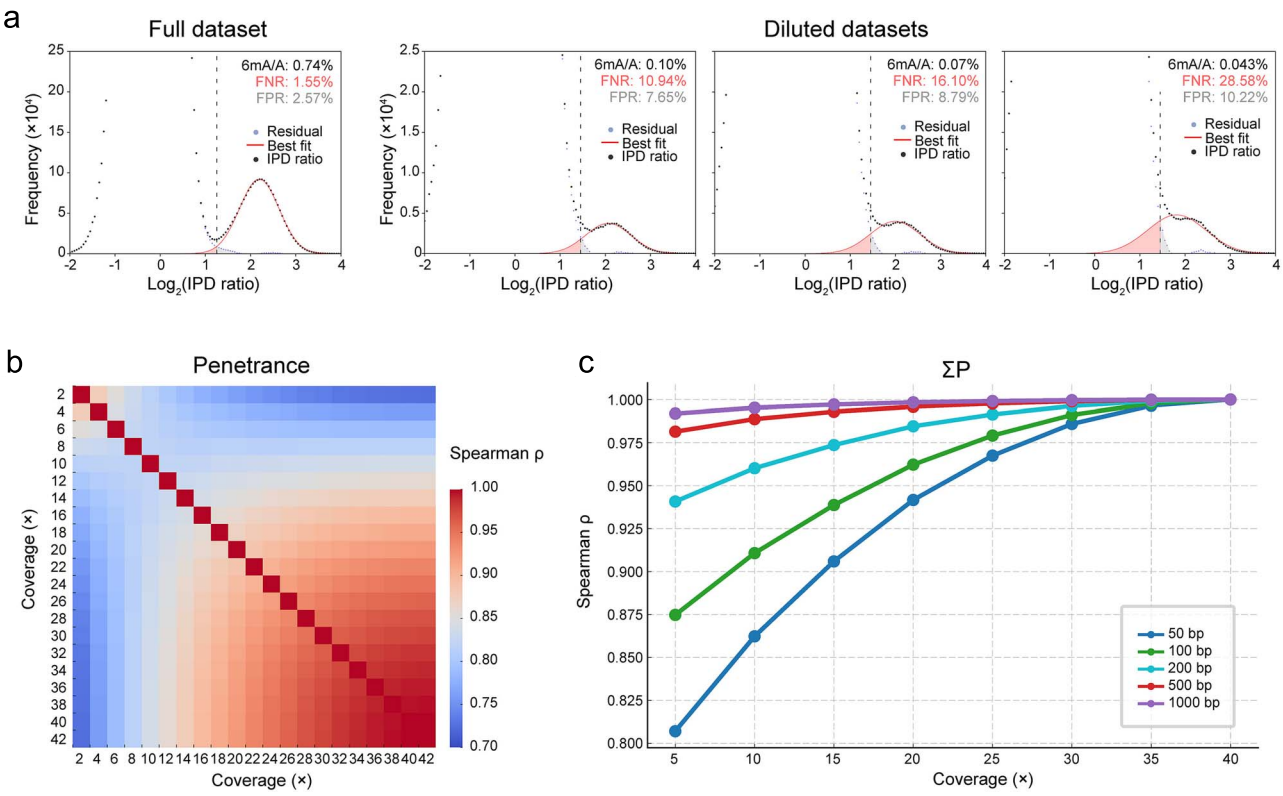

Figure 5. Impact of 6mA levels and sequencing coverage on reliable 6mA detection. (a) Deconvolution of the 6mA peak and unmodified A peak for IPD ratio distribution in diluted datasets of native DNA sample. The native DNA data were mixed with WGA data with the ratios of data size at 1:6, 1:9, and 1:15. Residual is the difference between authentic IPD distribution and the best fit curve. The IPD ratio cutoff was marked by dashed line. (b) Correlation of penetrance in gradient-partitioned CCS datasets of *Escherichia coli* at varying coverages. The full 42× dataset was subdivided into 21 partial datasets, each with 2× decremental coverage. Spearman correlation coefficient was calculated between each pair of datasets. (c) Correlation of ΣP (the sum of penetrance within 50, 100, 200, 500, and 1000 bp of the reference genome) in gradient-partitioned CCS datasets of *E. coli* with different coverages. The full 42× dataset was partitioned into 8 partial datasets with 5× descending per dataset. Spearman correlation coefficient was calculated between each pair of datasets.

At the single-molecule level, SMAC can provide reliable 6mA detection without requiring high sequencing coverage. However, at the genome-wide level, a certain sequencing depth is necessary. We performed a gradient partition of an *E. coli* CCS dataset with approximately 42× depth and used SMAC to calculate the penetrance of individual A sites and the 6mA level of a specific region (ΣP, the sum of penetrance within a specific region of the reference genome). The correlation of penetrance between partial dataset and full dataset gradually increased and stabilized as the coverage of the partial dataset approached that of the full dataset, with the Spearman correlation coefficient surpassing 0.9 once the depth exceeded 20× (Fig. 5b). A similar pattern is observed with ΣP (Fig. 5c). Additionally, higher sequencing depth could reduce the discrepancy on penetrance and methylation level between SMAC and the standard SMRT-seq pipeline (Supplementary Fig. S5). Therefore, SMAC recommends a minimal sequencing depth of over 20× coverage for genome-wide 6mA assessment.

Conclusion and perspective

In this work, we developed SMAC, an automated framework for detecting 6mA based on SMRT CCS data. SMAC applies rigorous data preprocessing with adjustable parameters, allowing users to balance data yield and quality according to their specific needs. By analyzing molecule-specific IPD ratios and leveraging the statistical distribution characteristics of these ratios, SMAC identifies 6mA at single-molecule, single-nucleotide resolution with high confidence, delivering results that are straightforward to visualize. It detects low penetrance sites and uncovers hidden 6mA patterns at 6mApT dinucleotides that may be overlooked or misinterpreted in studies using the standard SMRT-seq pipeline. Apart from the standard SMRT-seq pipeline, several methods have been developed to estimate 6mA profiles based on CCS data. Compared with existing analysis processes, SMAC stands out in data preprocessing, evaluation, resolution, and user experience (Supplementary Table S3).

In addition to software tools that rely on SMRT CCS data for identifying 6mA sites, another category of tools has been developed to predict 6mA sites without requiring sequencing, such as 6mA-Finder, csDMA, and PSAC-6mA [38–40]. These tools utilize deep learning models trained on existing databases, enabling them to predict 6mA sites with reasonable accuracy based solely on sequence features of the target species. Such tools are highly user-friendly and require minimal experimental and computational resources, making them particularly suitable for rapid and large-scale identification of potential 6mA sites across diverse species. However, these methods have notable limitations. They are unable to reveal the heterogeneity among individual DNA molecules due to insufficient characterization at the single-molecule resolution, relying instead on consensus sequence features. Most critically, these sequence-based prediction tools cannot track the dynamic changes in 6mA modifications across different physiological states, tissues, or cell lines within the same

species, as the underlying genomic sequences remain identical (Supplementary Table S3).

The demarcation of 6mA states by SMAC provides the resolution needed to study the catalytic features of 6mA methyltransferase (MTase) [41]. SMAC is essential in revealing that the catalytic product of the *de novo* MTase AMT2 is hemi-6mApT [42]. Additionally, by tracing cell cycle-dependent 6mA dynamics, SMAC uncovers the hemi-to-full conversion mediated by the maintenance MTase AMT1 [10]. These examples demonstrate the power of SMRT CCS and SMAC, which are beyond the reach of traditional CLR-based ensemble analysis [7].

SMAC can also empower us to explore 6mA functions during environmental responses. In the ciliate *Pseudocohnilembus persalinus* and the fungi *Phycomyces blakesleeanus and Mucor lusitanicus*, symmetric and asymmetric methylated 6mApT have been demonstrated to fulfill different biological roles [12, 32]. However, both studies relied on the CLR data analyzed by the standard SMRT-seq pipeline [12, 32]. It is conceivable that SMAC could offer more precise determination of hemi and full methylation in *P. persalinus*, *P. blakesleeanus*, *M. lusitanicus* and other species, thereby enhancing our understanding of the biological significance of differentially methylated 6mA.

The PacBio Sequel II system offers significantly higher data output compared to its predecessor, the RS II system, making it the platform of choice in most studies utilizing PacBio sequencing. SMAC is specifically designed to process subreads data generated from the Sequel II system. However, with the advancements in PacBio's sequencing technology, the recently introduced Revio system delivers even greater data yields than Sequel II. As a result, future adjustment and optimization to our workflow may be necessary to accommodate the increased data capacity of the Revio system.

Key Points

- Single-molecule 6mA analysis of CCS reads (SMAC) is a user-friendly streamlined toolkit designed for detecting 6mA based on SMRT circular consensus sequencing (CCS) data requiring only raw sequencing data input to generate quality control information, statistical analysis, and visualized results.
- SMAC employs comprehensive data pretreatment to minimize background noise, thereby enhancing the accuracy and reliability of the results.
- SMAC outperforms the standard SMRT-seq pipeline in identifying genome-wide 6mA sites, offering superior sensitivity at the single-molecule level.
- A Gaussian distribution fitting approach is utilized to objectively define the cutoff for 6mA site detection, ensuring reliable results.
- SMAC provides higher precision in detecting 6mA within ApT (a short DNA sequence composed of an adenine followed by a thymine, connect by a phosphodiester bond) dinucleotides, accurately distinguishing between full-methylated and hemi-methylated states.

Acknowledgments

High-performance computing resources for data processing were provided by the Institute of Evolution and Marine Biodiversity, High-Performance Biological Supercomputing Center, and Marine Big Data Center of Institute for Advanced Ocean Study at the Ocean University of China (OUC), as well as the Center for Advanced Research Computing (CARC) at the University of Southern California. The authors would like to thank Dr. Wentao Yang (University of Southern California, USA) for fruitful discussion. Our special thanks are given to Weibo Song (OUC) for his helpful suggestions during drafting the manuscript.

Author contributions

Haicheng Li, Junhua Niu, and Yalan Sheng (Methodology, Formal Analysis, Investigation, Writing—original draft, writing—review & editing) and Yifan Liu and Shan Gao (Conceptualization, Supervision, Funding Acquisition, Writing—review & editing).

Conflict of interest: None declared.

Funding

This work was supported by the National Natural Science Foundation of China (32125006 to S.G.), Natural Science Foundation of Shandong Province of China (ZR2024ZD40 to S.G.), the Science & Technology Innovation Project of Laoshan Laboratory (LSKJ202203203 to S.G.), and the National Science Foundation (MCB-2435178 to Y.L.).

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Data availability

CCS datasets were downloaded from the NCBI database with the following BioProject accession numbers: *Tetrahymena thermophila* (PRJNA932808), *Caenorhabditis elegans* (PRJNA857919), and various other organisms including human PBMC (peripheral blood mononuclear cells), *Arabidopsis thaliana*, *Drosophila melanogaster*, *Escherichia coli*, *Chlamydomonas reinhardtii*, and a mixed sample containing *Escherichia coli*, *Helicobacter pylori*, and *Saccharomyces cerevisiae* (PRJNA667898). The SMAC toolkit is available along with a tutorial at GitHub (https://github.com/liiihc/SMAC).

References

1. Fu Y, Luo G-Z, Chen K. *et al.* N^6-methyldeoxyadenosine marks active transcription start sites in *Chlamydomonas*. *Cell* 2015;**161**:879–92. https://doi.org/10.1016/j.cell.2015.04.010.
2. Zhang GQ, Huang H, Liu D. *et al.* N^6-methyladenine DNA modification in *Drosophila*. *Cell* 2015;**161**:893–906. https://doi.org/10.1016/j.cell.2015.04.018.
3. Wu TP, Wang T, Seetin MG. *et al.* DNA methylation on N^6-adenine in mammalian embryonic stem cells. *Nature* 2016;**532**:329–33. https://doi.org/10.1038/nature17640.
4. Wang Y, Chen X, Sheng Y. *et al.* N^6-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. https://doi.org/10.1093/nar/gkx883.
5. Xie Q, Wu TP, Gimple RC. *et al.* N^6-methyladenine DNA modification in glioblastoma. *Cell* 2018;**175**:1228–1243.e1220. https://doi.org/10.1016/j.cell.2018.10.006.

6. Beh LY, Debelouchina GT, Clay DM. *et al.* Identification of a DNA N^6-adenine methyltransferase complex and its impact on chromatin organization. *Cell* 2019;**177**:1781–1796.e1725. https://doi.org/10.1016/j.cell.2019.04.028.
7. Wang Y, Sheng Y, Liu Y. *et al.* A distinct class of eukaryotic MT-A70 methyltransferases maintain symmetric DNA N^6-adenine methylation at the ApT dinucleotides as an epigenetic mark associated with transcription. *Nucleic Acids Res* 2019;**47**:11771–89. https://doi.org/10.1093/nar/gkz1053.
8. Li Z, Zhao S, Nelakanti RV. *et al.* N^6-methyladenine in DNA antagonizes SATB1 in early development. *Nature* 2020;**583**:625–30. https://doi.org/10.1038/s41586-020-2500-9.
9. Zhang S, Li B, Du K. *et al.* Epigenetically modified N^6-methyladenine inhibits DNA replication by human DNA polymerase iota. *Biochimie* 2020;**168**:134–43. https://doi.org/10.1016/j.biochi.2019.10.018.
10. Sheng Y, Wang Y, Yang W. *et al.* Semiconservative transmission of DNA N^6-adenine methylation in a unicellular eukaryote. *Genome Res* 2024;**34**:740–56. https://doi.org/10.1101/gr.277843.123.
11. Sheng Y, Pan B, Wei F. *et al.* Case study of the response of N^6-methyladenine DNA modification to environmental stressors in the unicellular eukaryote *Tetrahymena thermophila*. *mSphere* 2021;**6**:e01208–20. https://doi.org/10.1128/mSphere.01208-20.
12. Liu Y, Niu J, Ye F. *et al.* Dynamic DNA N^6-adenine methylation (6mA) governs the encystment process, showcased in the unicellular eukaryote *Pseudocohnilembus persalinu*. *Genome Res* 2024;**34**:256–71. https://doi.org/10.1101/gr.278796.123.
13. Boulias K, Greer EL. Means, mechanisms and consequences of adenine methylation in DNA. *Nat Rev Genet* 2022;**23**:411–28. https://doi.org/10.1038/s41576-022-00456-x.
14. Feng X, He C. Mammalian DNA N^6-methyladenosine: challenges and new insights. *Mol Cell* 2023;**83**:343–51. https://doi.org/10.1016/j.molcel.2023.01.005.
15. Lyu C, Wang H-D, Lai W. *et al.* Identification and quantification of DNA N^6-methyladenine modification in mammals: a challenge to modern analytical technologies. *Curr Opin Chem Biol* 2023;**73**:102259. https://doi.org/10.1016/j.cbpa.2022.102259.
16. Eid J, Fehr A, Gray J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* 2009;**323**:133–8. https://doi.org/10.1126/science.1162986.
17. Flusberg BA, Webster DR, Lee JH. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* 2010;**7**:461–5. https://doi.org/10.1038/nmeth.1459.
18. Fang G, Munera D, Friedman DI. *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. https://doi.org/10.1038/nbt.2432.
19. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. *Genome Biol* 2013;**14**:405. https://doi.org/10.1186/gb-2013-14-6-405.
20. Schadt EE, Banerjee O, Fang G. *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. https://doi.org/10.1101/gr.136739.111.
21. Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* 2015;**13**:278–89. https://doi.org/10.1016/j.gpb.2015.08.002.
22. Ardui S, Ameur A, Vermeesch JR. *et al.* Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res* 2018;**46**:2159–68. https://doi.org/10.1093/nar/gky066.
23. Beaulaurier J, Zhang X-S, Zhu S. *et al.* Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. *Nat Commun* 2015;**6**:7438. https://doi.org/10.1038/ncomms8438.
24. Abdulhay NJ, McNally CP, Hsieh LJ. *et al.* Massively multiplex single-molecule oligonucleosome footprinting. *Elife* 2020;**9**:e59404. https://doi.org/10.7554/eLife.59404.
25. Stergachis AB, Debo BM, Haugen E. *et al.* Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* 2020;**368**:1449–54. https://doi.org/10.1126/science.aaz1646.
26. Feng Z, Fang G, Korlach J. *et al.* Detecting DNA modifications from SMRT sequencing data by Modeling sequence context dependence of polymerase kinetic. *PLoS Comput Biol* 2013;**9**:e1002935. https://doi.org/10.1371/journal.pcbi.1002935.
27. Kong Y, Cao L, Deikus G. *et al.* Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. *Science* 2022;**375**:515–22. https://doi.org/10.1126/science.abe7489.
28. Zhang J, Peng Q, Ma C. *et al.* 6mA-sniper: quantifying 6mA sites in eukaryotes at single-nucleotide resolution. *Science Advances* 2023;**9**:eadh7912. https://doi.org/10.1126/sciadv.adh7912.
29. Liang Z, Shen L, Cui X. *et al.* DNA N^6-adenine methylation in *Arabidopsis thaliana*. *Dev Cell* 2018;**45**:406–416.e403. https://doi.org/10.1016/j.devcel.2018.03.012.
30. Zhang Q, Liang Z, Cui X. *et al.* N^6-methyladenine DNA methylation in *japonica* and *indica* rice genomes and its association with gene expression, plant development, and stress responses. *Mol Plant* 2018;**11**:1492–508. https://doi.org/10.1016/j.molp.2018.11.005.
31. Zhou C, Wang C, Liu H. *et al.* Identification and analysis of adenine N^6-methylation sites in the rice genome. *Nat Plants* 2018;**4**:554–63. https://doi.org/10.1038/s41477-018-0214-x.
32. Lax C, Mondo SJ, Osorio-Concepción M. *et al.* Symmetric and asymmetric DNA N^6-adenine methylation regulates different biological responses in Mucorales. *Nat Commun* 2024;**15**:6066. https://doi.org/10.1038/s41467-024-50365-2.
33. Hahn A, Hung GCC, Ahier A. *et al.* Misregulation of mitochondrial 6mA promotes the propagation of mutant mtDNA and causes aging in *C. elegans*. *Cell Metab* 2024;**36**:2528–2541.e2511. https://doi.org/10.1016/j.cmet.2024.07.020.
34. Zhao L, Gao F, Gao S. *et al.* Biodiversity-based development and evolution: the emerging research systems in model and non-model organisms. *Sci China Life Sci* 2021;**64**:1236–80. https://doi.org/10.1007/s11427-020-1915-y.
35. Ye F, Chen X, Li Y. *et al.* Comprehensive genome annotation of the model ciliate *Tetrahymena thermophila* by in-depth epigenetic and transcriptomic profiling. *Nucleic Acids Res* 2024:**53**;gkae1177. https://doi.org/10.1093/nar/gkae1177.
36. Pan B, Ye F, Li T. *et al.* Potential role of N^6-adenine DNA methylation in alternative splicing and endosymbiosis in *Paramecium bursaria*. *iScience* 2023;**26**:106676. https://doi.org/10.1016/j.isci.2023.106676.
37. Zhao H, Ma J, Tang Y. *et al.* Genome-wide DNA N^6-methyladenosine in *Aeromonas veronii* and *Helicobacter pylori*. *BMC Genomics* 2024;**25**:161. https://doi.org/10.1186/s12864-024-10074-y.
38. Xu H, Hu R, Jia P. *et al.* 6mA-finder: a novel online tool for predicting DNA N^6-methyladenine sites in genomes. *Bioinformatics* 2020;**36**:3257–9. https://doi.org/10.1093/bioinformatics/btaa113.

39. Liu Z, Dong W, Jiang W. *et al.* csDMA: an improved bioinformatics tool for identifying DNA 6mA modifications via Chou's 5-step rule. *Sci Rep* 2019;**9**:13109. https://doi.org/10.1038/s41598-019-49430-4.
40. Zhou Z, Xiao C, Yin J. *et al.* PSAC-6mA: 6mA site identifier using self-attention capsule network based on sequence-positioning. *Comput Biol Med* 2024;**171**:108129. https://doi.org/10.1016/j.compbiomed.2024.108129.
41. Wang Y, Nan B, Ye F. *et al.* Dual modes of DNA N^6-methyladenine maintenance by distinct methyltransferase complexes. *Proc Natl Acad Sci* 2025;**122**:e2413037121. https://doi.org/10.1073/pnas.2413037121.
42. Cheng T, Zhang J, Li H. *et al.* Identification and characterization of the *de novo* methyltransferases for eukaryotic N^6-methyladenine (6mA). *bioRxiv* 2024;**2024**:2003.2025.586193.

Briefings in Bioinformatics, 2025, **26(2)**, bbaf153
https://doi.org/10.1093/bib/bbaf153
Problem Solving Protocol