Unimeth: A unified transformer framework for accurate DNA methylation detection from nanopore reads

Shaokai Wang[1,2,3], Yifu Xiao[1,2,3], Tao Sheng[1,2,3], Neng Huang[4,5], Yi Shu[6], Jixian Zhai[6], Feng Luo[7], Peng Ni[1,2,3,*]

[1]School of Computer Science and Engineering, Central South University, Changsha 410083, China

[2]Xiangjiang Laboratory, Changsha 410205, China

[3]Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha 410083, China

[4]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA

[5]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

[6]Department of Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen, China

[7]School of Computing, Clemson University, Clemson, SC, 29634-0974, USA

*To whom correspondence should be addressed:

Peng Ni. Email: nipeng@csu.edu.cn;

Abstract

Nanopore sequencing has emerged as a powerful technology for DNA methylation detection, particularly in repetitive genomic regions and at the haplotype scale. However, existing computational methods show inconsistent accuracy across sequence contexts, species, and sequencing chemistries. Here, we present Unimeth, a unified transformer-based framework that simultaneously predicts multi-site methylation from nanopore reads. Unimeth employs a patch-based architecture and a three-phase training strategy, including pre-training, read-level fine-tuning, and site-level calibration, to fully leverage genome-wide methylation information. In comprehensive benchmarks involving 20 samples spanning 13 species, Unimeth consistently outperforms state-of-the-art methods. This unified approach demonstrates superior accuracy and significantly reduced false positives across a wide range of scenarios, including the detection of both 5mC and 6mA, application in organisms from mammals and plants to bacteria, analysis of both wild-type and mutant samples, and use of both R10.4 and R9.4 pore chemistries. Furthermore, Unimeth is demonstrated to be a highly accurate tool for methylation analysis in transposons and centromeric regions.

Keywords: nanopore sequencing, methylation, deep learning

Introduction

DNA methylation, such as 5-methylcytosine (5mC), N^6-methyladenine (6mA), and N^4-methylcytosine (4mC), is essential for various biological processes in bacteria, mammals, and plants[1-3]. Traditionally, DNA methylation has been extensively studied using microarray or short-read sequencing-based methods with specialized library preparation, such as chemical conversion, antibody enrichment, and enzyme digestion[4,5]. However, these methods have limitations and may introduce biases due to their pretreatment of genomic DNA[5]. For example, methylated DNA immunoprecipitation-sequencing (MeDIP-seq), which uses antibodies directed against 5mC, lacks single-nucleotide resolution and is biased toward highly methylated regions[6]. The bisulfite conversion in BS-seq is usually not 100% efficient and can lead to DNA degradation, further resulting in non-uniform sequencing coverage and false positive calls[4,5,7,8].

In recent years, long-read sequencing technologies, including Oxford Nanopore Technologies (ONT) nanopore sequencing and PacBio single-molecule real-time (SMRT) sequencing, have increasingly been utilized for DNA methylation detection[9,10]. Long-read sequencing offers several advantages over short-read sequencing. First, long-read sequencing can sequence native DNA fragments, enabling the direct detection of DNA methylation without extra laboratory techniques. This feature further leads to more uniform sequencing coverage and allows the simultaneous detection of multiple methylation types at the single-molecule, single-nucleotide level[5,11]. Second, the long-read lengths enable the epigenetic profiling of highly repetitive regions (such as transposons[12] and centromeres[13,14]), and across entire haplotype blocks[15,16].

Nanopore sequencing has emerged as a highly promising technology for profiling DNA methylation[17]. Numerous computational methods have been proposed for DNA methylation

detection using nanopore sequencing[18-20]. These methods mainly fall into three categories: (1) statistics-based methods that apply statistical testing comparing nanopore current signals of native DNA with an unmodified control, such as Tombo[21]; (2) model-based methods that use pretrained machine-learning or deep-learning models to classify the methylation status of target sites, such as nanopolish[22] and DeepSignal[23]; (3) basecalling-based methods that directly identify modified bases during basecalling with an extended alphabet, such as Taiyaki[24]. Especially, with the release of the latest R10.4.1 kits by ONT, several tools have been developed or updated to detect methylation from R10.4.1 reads, such as DeepMod2[25], Rockfish[26], and DeepPlant[27].

Despite significant advancements, methods for detecting DNA methylation using nanopore sequencing remain limited in several aspects, including generalizability across diverse species and sequence contexts, computational efficiency, and standardization of input/output formats. mCaller, a pioneering tool designed for DNA 6mA detection in bacteria, is restricted to R9.4.1 chemistry[28]. The official ONT basecaller, Dorado, which is actively maintained and frequently updated by ONT, supports the detection of various DNA modifications, including 5mC, 4mC, 6mA, and 5-hydroxymethylcytosine (5hmC) during basecalling[29]. However, neither the training data nor the training strategy of Dorado is publicly available. Furthermore, the performance of Dorado for 5mC detection has yet to be thoroughly validated, particularly across different species. DeepMod2[25] and Rockfish[26] are the two state-of-the-art third-party methods for 5mC detection in human and mouse cell lines using R10.4.1 chemistry. On the human NA12878 R10.4.1 dataset, Rockfish achieves an over 0.97 read-level accuracy for 5mC detection. However, both Rockfish and DeepMod2 are currently restricted to CpG sites and lack the capability to detect 5mC in CHG and CHH these non-CpG contexts (CHG and CHH, H represents A, C, or T), which are crucial for stably

transposon silencing and gene expression regulation. Nevertheless, this task presents unique challenges because methylated sites constitute only a small fraction of all sites, resulting in a severe class imbalance. This imbalance demands models with exceptional sensitivity to reliably identify rare methylation events. Existing approaches have primarily addressed this through dataset modification strategies for training. DeepSignal-plant[30] employed semi-supervised learning to generate pseudo-labels for balancing training samples. DeepPlant[27] further incorporated training samples from diverse species with elevated cutoff of methylation frequencies.

While existing methods achieve reliable performance in human or cross-species methylation detection, fundamental limitations persist in their detection frameworks. Although various computational architectures have been explored in recent methods, they remain constrained by the k-mer-based paradigm, which predicts methylation at individual sites by analyzing localized k-mers and their corresponding signal segments. This limitation not only prevents the model from obtaining contextual biological information by failing to capture long-range interdependencies between neighboring methylation events, but also reduces efficiency due to computational redundancy. Moreover, most existing tools rely on read-level binary labels for training, utilizing only either fully methylated or unmethylated sites. As a result, when confronted with partially methylated sites, these methods may lead to bias.

To address the limitations of current methods, we propose Unimeth, a unified transformer-based framework that simultaneously detects multiple DNA methylation sites from nanopore reads. First, we replace the conventional *k*-mer-based prediction paradigm with a patch-based framework, which processes continuous segments of reads by incorporating both raw signals and corresponding base sequences, and simultaneously predicts the methylation status of all targeted sites within each patch.

This design leverages attention mechanisms to explicitly model dependencies among neighboring sites, enabling more biologically informed and context-aware methylation detection. Second, beyond conventional read-level training, Unimeth incorporates a site-level training phase that directly utilizes continuous methylation ratios from BS-seq as training labels. Using predictions from the fine-tuned model and empirical methylation levels, we reassign confident positive and negative labels at the site level, enabling the model to learn from a more representative and continuous spectrum of methylation states, including partially methylated sites often ignored by other methods.

These innovations together lead to substantial improvements in the accuracy and robustness of 5mC and 6mA detection. We trained separate Unimeth models for nanopore reads generated using R10.4.1 5 kHz, R10.4.1 4 kHz, and R9.4.1 chemistries. We then comprehensively benchmarked Unimeth using sequencing data of 13 species (including 10 plants, 2 mammals, and 1 bacterium) from publicly available datasets and in-house resources. Especially, we generated nanopore reads of mutant strains of *A. thaliana* to benchmark the false positive rates of Unimeth and other methods on 5mC detection. According to the benchmarking, Unimeth achieves consistently higher performance in genome-wide methylation detection across multiple species and cell lines. For 5mC detection in plants, Unimeth exhibits exceptional capability in non-CpG sequence contexts, where severe class imbalance typically impedes conventional methods. For CpG-only 5mC detection, our model achieves the highest accuracy across multiple human cell lines as well as in cross-species validation. For 6mA detection, our model also demonstrates strong performance by reducing false positives, leading to the discovery of more accurate positions of nucleosomes by using Fiber-seq[31]. In conclusion, by using nanopore sequencing, Unimeth enables accurate genome-wide andstrand-specific methylation detection at single-nucleotide resolution. We further packaged

Unimeth as an easy-to-use tool that supports model pre-training, fine-tuning, and methylation detection, with standardized input formats (POD5/SLOW5 and BAM) and output formats (ModBam and BED).

Results

The framework of Unimeth

Unimeth predicts methylation status of all targeted sites with a given nanopore read (Fig. 1). Unimeth divides nanopore read signals and their corresponding basecalled DNA sequences into patches. The Unimeth framework uses an encoder-decoder framework based on transformer modules. The normalized signals are preprocessed by a 1-D convolutional module for extracting signal features and then used as the input of the encoder modules. On the basecalled read sequences, special tokens indicating motif types are inserted after the corresponding cytosine tokens. Unimeth employs a 12-layer transformer architecture for both its encoder and decoder. The model predicts methylation status by outputting one of two special tokens: [-] for unmethylated or [+] for methylated. All methylation sites within a patch are detected simultaneously.

To train Unimeth, we employ a multi-phase training strategy consisting of pre-training, read-level fine-tuning, and site-level calibration (Fig. 1a, Methods). During the pre-training phase, the model learns basecalling by taking raw nanopore signals as input and generating the corresponding nucleotide sequences. In the fine-tuning phase, high-confidence methylated and unmethylated sites are selected to fine-tune the model for methylation prediction by introducing special tokens into the target labels. In the calibration phase, genome-wide methylation information, including partially methylated sites, is utilized to further refine the model. This multi-phase training strategy enhances

the ability of Unimeth to accurately predict uncertain and partially methylated sites.

We have packaged Unimeth as an easy-to-use tool for genome-wide methylation detection from nanopore reads (Fig. 1b). Unimeth takes raw signals in POD5 or SLOW5 format and basecalled reads in BAM format as input, and outputs per-read, single-nucleotide methylation results in modBAM or text format. Unimeth can also extract per-site methylation frequencies from aligned reads in BAM format for further downstream analysis.

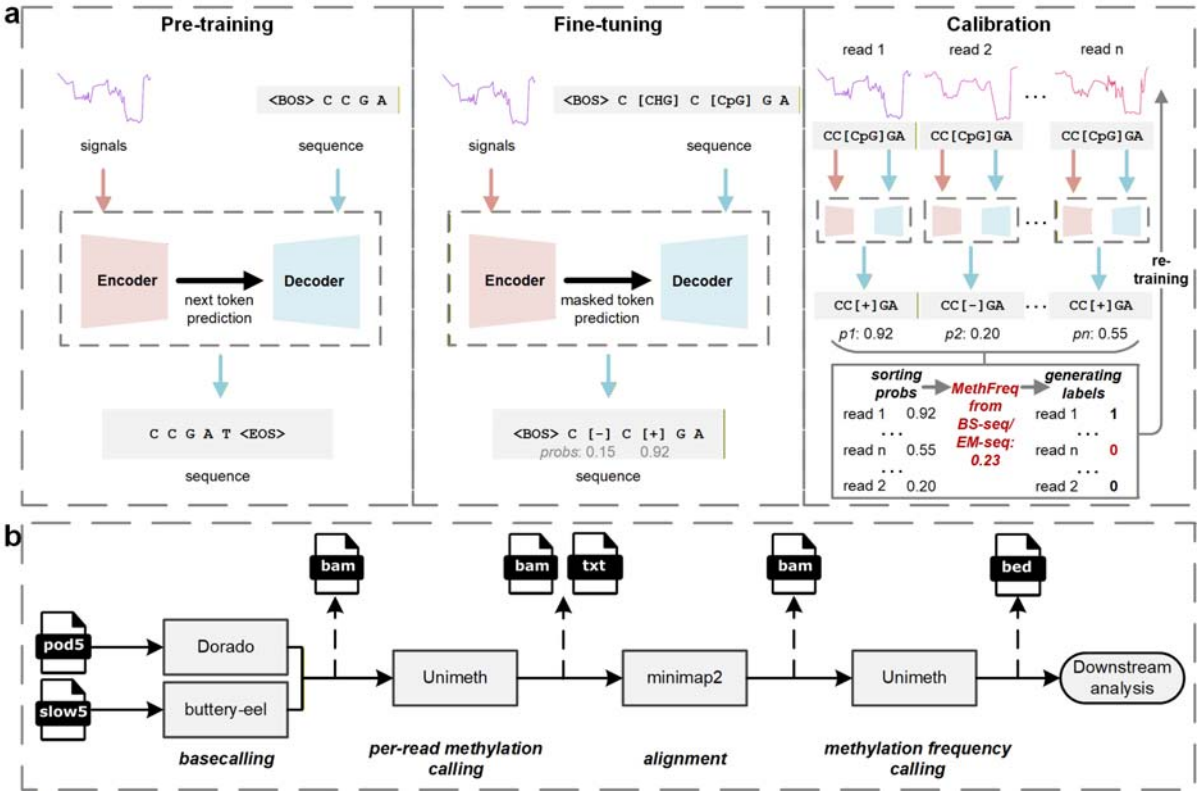

Fig. 1 The framework of Unimeth. **a** The training process of Unimeth (probs: probabilities, MethFreq: methylation frequency). Nanopore raw signals are segmented and aligned to corresponding 256-bp sequence patches. Special context tokens (CpG, CHG, CHH) are inserted to mark targeted sites. The encoder processes normalized signal embeddings, while the decoder integrates sequence and context information to jointly call all methylation sites within a patch. The three subfigures illustrate the three-phase training workflow. During pre-training, the model learns signal-to-base translation. Fine-tuning incorporates read-level methylation labels to discriminate fully methylated/unmethylated sites. Site-level calibration refines labels for partially methylated sites using BS-seq/EM-seq methylation frequencies. **b** Typical workflow of methylation calling using Unimeth.

Read-level evaluation of 5mC detection in plants

For 5mC detection in plants, we trained two separate models for R10.4.1 5 kHz and R9.4.1 chemistries. For R10.4.1 5 kHz kits, we first collected nanopore reads of nine plant species[27] including Arabidopsis thaliana (*A. thaliana*), Beta vulgaris (*B. vulgaris*), Citrus sinensis (*C. sinensis*), Oryza sativa (*O. sativa*), Ricinus communis (*R. communis*), Salvia miltiorrhiza (*S. miltiorrhiza*), Solanum lycopersicum (*S. lycopersicum*), Solanum tuberosum (*S. tuberosum*), and Vitis vinifera (*V. vinifera*).

The methylation levels in reads varied significantly across these species (Supplementary Data 2). For instance, methylation levels of the CpG motif ranged from 26.28% in *A. thaliana* to 80.60% in *S. lycopersicum*. Methylation levels within the same species also differs substantially among motifs. For instance, *B. vulgaris* exhibited a CpG methylation level of 80.12%, while its CHG and CHH methylation levels are 52.28% and 9.66%, respectively. These pronounced interspecies and motif-specific variations underscore the challenge of developing a method adaptable to diverse species and motifs.

We evaluated Unimeth against Rockfish, DeepMod2, Dorado, and DeepPlant across the nine datasets. To eliminate threshold-related biases, we used the area under the curve (AUC) as the evaluation metric. The results (Fig. 2) demonstrate that Unimeth consistently outperformed Rockfish and DeepMod2, two CpG-specific models, in CpG methylation detection across all plant species. Notably, while they represent state-of-the-art methods for CpG detection in human data, their performance was suboptimal in plant species. This highlights Unimeth's superior generalization capability across diverse biological contexts. Dorado and DeepPlant support the detection of all 5mC methylation types, including CpG, CHG, and CHH. Unimeth achieves the highest AUC in 8 out of 9

datasets for both CpG and CHG methylation, with *S. tuberosum* being the sole exception. Across these datasets, Unimeth attains an average AUC of 98.60% for CpG and 99.09% for CHG, surpassing the previous state-of-the-art method, DeepPlant, which achieves 98.26% for CpG and 98.20% for CHG, respectively. For CHH detection where methylation levels are typically lower than 1%, Unimeth demonstrates even more pronounced superiority, achieving an AUC of 98.23%, significantly higher than DeepPlant's 92.54%. To further evaluate performance under extremely unbalanced methylation ratios in the CHH context, we assessed the methods using precision, recall, and the area under the precision-recall curve (AUPR). Unimeth achieved a precision of 26.08%, a recall of 80.02%, and an AUPR of 50.45%, representing a substantial improvement over DeepPlant, which yielded a precision of 17.14%, a recall of 73.16%, and an AUPR of 27.09% (Supplementary Fig. 1).

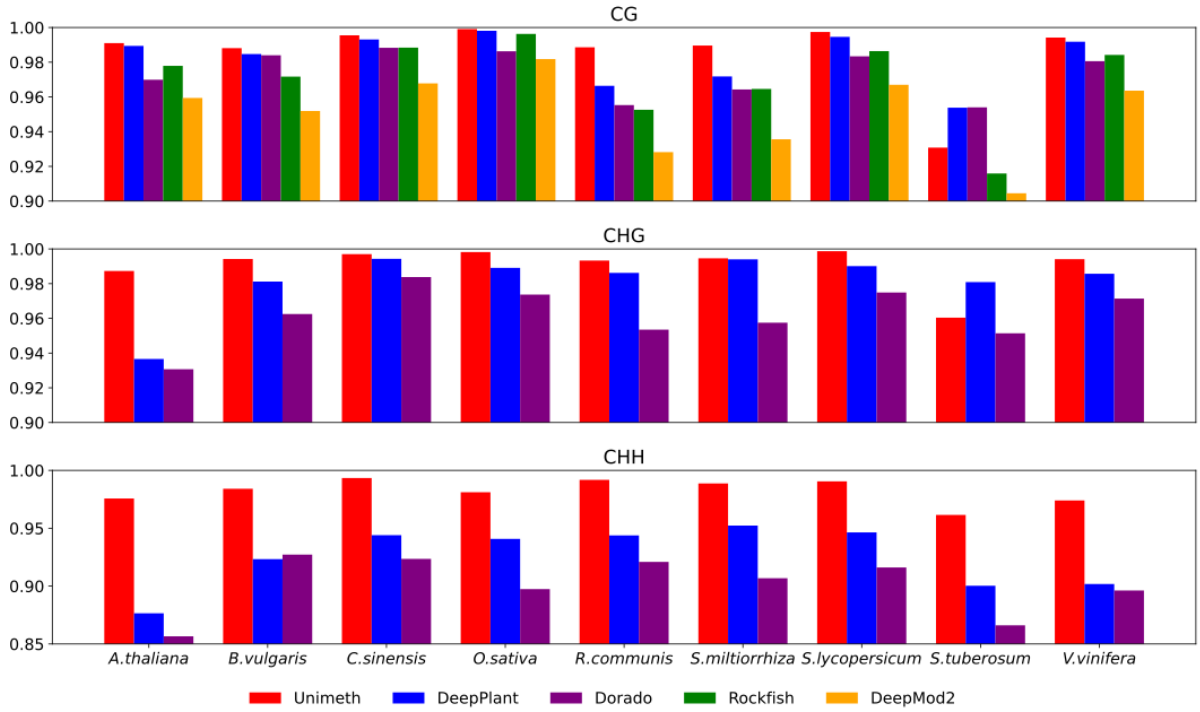

Fig. 2 Read-level performance comparison of five prediction methods (Unimeth, DeepPlant, Dorado, Rockfish, DeepMod2) across the three methylation motifs, with AUC values plotted for each method and motif.

This enhancement at the read level is primarily attributed to our success in mitigating the high false positive rate (FPR) observed in other methods. We calculated the FPR across the nine species. As shown in Supplementary Fig. 2, our method substantially reduces the FPR, particularly at non-CpG contexts. Specifically, on CHG sites, Unimeth lowered the average FPR to 0.63%, compared to 0.91% for DeepPlant. On CHH sites, Unimeth achieved an even more dramatic reduction, bringing the FPR down to 0.7% from 1.42%. This reduction in false positives stems from our training strategy. Instead of artificially balancing the dataset by adding positive samples for the underrepresented non-CpG contexts, we trained the model on data with its original ratios. This approach prevented the model from being biased towards the positive class, thereby effectively resolving the issue of high false positives.

To assess the generalizability and robustness of our model against cross-species and batch effects, we benchmarked Unimeth against DeepPlant and Dorado on three in-house datasets. These datasets encompassed *O. sativa* and *A. thaliana*, which contained potential batch variations, as well as the completely unseen *Musa* spp., a species absent from prior training data. Performance was compared across CpG, CHG, and CHH contexts. The AUC values for all three methods are summarized in Supplementary Fig. 3a. Despite these challenges, Unimeth consistently outperformed both competitors, achieving the highest AUC and demonstrating strong robustness and generalization capability. Compared with the previous best-performing method DeepPlant, Unimeth improved by 2.04% and 5.14% on CHG and CHH contexts.

Despite being an older chemistry, the R9.4.1 platform remains widely used. To support the scientific community in maximizing the value of these existing datasets, we also trained a model for 5mC detection using R9.4.1 chemistry, maintaining the same architecture as our R10.4.1 model. To

evaluate performance, we collected *A. thaliana* and *O. sativa* to compared Unimeth with DeepSignal-plant[30], the best method for 5mC methylation detection tool on R9.4.1 chemistry. At the read level, Unimeth demonstrates superior performance across all contexts (Supplementary Fig. 4a and b). In terms of AUPR, it achieves average improvements of 4.75% for CpG, 10.02% for CHG, and 43.42% for CHH over DeepSignal-plant, highlighting its robustness in detecting methylation across different platforms.

To further validate our method's capability in reducing false positives, we leveraged nanopore data of the *met1-3* and *ddcc* mutants of *A. thaliana*[32]. The *met1-3* mutant is characterized by a loss of CpG methylation, while *ddcc* lacks methylation in non-CpG contexts. Compared to DeepSignal-plant, Unimeth demonstrated a dramatic reduction in detected methylation in these mutants, which serves as definitive evidence for its lower false positive rate (Supplementary Fig. 5). In the *met1-3* data, the methylation level in CpG contexts detected by DeepSignal-plant dropped from approximately 27% in the wild-type to 13%, while Unimeth reduced to just 4%. Similarly, in the *ddcc* mutant, the methylation levels detected by DeepSignal-plant in CHG and CHH contexts were 2.843% and 2.554%, respectively. Unimeth further suppressed these false positives to 1.248% for CHG and 0.375% for CHH. These results of the mutant strains provide compelling evidence that Unimeth effectively minimizes false positive calls.

Site-level evaluation of 5mC detection in plants

In addition to read-level analysis, we evaluated model performance at the site level by aggregating methylation predictions across individual genomic sites. Site-level accuracy was quantified using Pearson correlation coefficients between model predictions and bisulfite sequencing (BS-seq) ground truth data across varying sequencing depths. Depth-specific correlations were calculated through

iterative random sampling of nanopore reads.

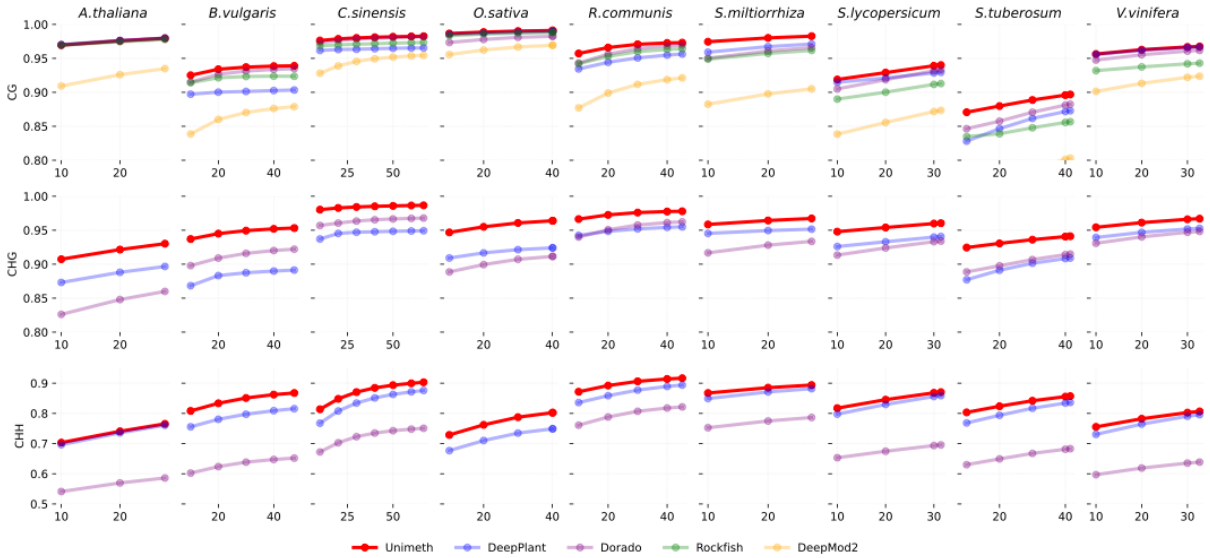

Fig. 3 Site-level performance of five prediction methods across varying sequencing depths. Each subplot illustrates the Pearson correlation with BS-seq results for different sequencing depths. The rows represent the three methylation motifs: CpG, CHG, and CHH, while the columns correspond to the nine species. The methods compared are Unimeth, DeepPlant, Dorado, Rockfish, and DeepMod2.

Fig. 3 presents the comparative performance of Unimeth, DeepPlant, Dorado, Rockfish, and DeepMod2 on the 9 datasets. For CpG detection, Unimeth outperformed CpG-specific methods such as Rockfish and DeepMod2 across varying sequencing depths. When subsampled to 10× depth, Unimeth achieved an average Pearson correlation of 0.948 across the nine species, surpassing Rockfish at 0.931 and DeepMod2 at 0.877. At full sequencing depth, with a mean of 39.98×, Rockfish and DeepMod2 improved to 0.945 and 0.907 respectively, yet still underperformed relative to Unimeth even at the lower 10× depth. Notably, Dorado and DeepPlant, which were trained on data beyond just human sources, showed stronger correlations: 0.936 and 0.934 at 10× depth, rising to 0.948 and 0.955 at full depth. Nevertheless, Unimeth achieved the highest overall correlation of 0.961, demonstrating its robustness across sequencing depth regimes and diverse species. While Unimeth did not achieve the best read-level performance for *S. tuberosum*, it demonstrated significantly superior performance at the site level: improved Dorado's 0.8464 to 0.8707 at 10×

depth. This discrepancy suggests that Unimeth's strength lies in its ability to aggregate and interpret methylation signals across multiple reads, particularly for species with complex methylation patterns like *S. tuberosum*.

For CHG and CHH detection, Unimeth demonstrates consistent superiority over existing methods across all nine species. In CHG detection, Unimeth achieves average Pearson correlations of 0.947 at 10× depth and 0.961 at full depth, significantly outperforming DeepPlant with values of 0.913 at 10× and 0.930 at full depth, as well as Dorado with 0.907 at 10× and 0.928 at full depth. The CHH detection task proves more challenging due to severe data imbalance, yet Unimeth maintains its advantage with correlations of 0.797 at 10× depth and 0.854 at full depth, surpassing DeepPlant which achieved 0.764 and 0.830 under the same conditions. This robust performance across both contexts highlights Unimeth's effectiveness in handling diverse methylation detection scenarios.

To fully assess our model on methylation detection, we further utilized in-house datasets with high sequencing depths of 82.46×, 99.2×, and 165.89× across three species (Supplementary Fig. 3b). For CpG detection, all methods performed well, with Unimeth achieving the highest correlations (0.9879, 0.9952, and 0.9727 at full depth). For CHG and CHH detection, Unimeth consistently outperformed both DeepPlant and Dorado across all depths. While performance trends approached the theoretical limit with increasing depth, Unimeth uniquely maintained its advantage: even at depths more than 80× where other methods plateaued, Unimeth’s correlations remained significantly higher. Unimeth also outperformed DeepSignal-plant on methylation detection from the R9.4.1 nanopore reads (Supplementary Fig. 4).

5mC profiling in repetitive genomic regions using Unimeth

With longer reads, nanopore sequencing has been demonstrated to profile methylation of more

cytosines than short-read sequencing[27,30], especially in repetitive genomic regions. By using the T2T reference genomes of *O. sativa* (T2T-NIP[33]), we explored how Unimeth performed in 5mC detection in all three contents in centromeres and TEs (transposonable elements). We used our in-house generated BS-seq and nanopore reads of *O. sativa* for the analysis, which have higher coverage than those from public datasets[27]. Unimeth profiles almost all of the cytosines (>99.9%) genome-widely in *O. sativa* by leveraging 99.2× nanopore reads (Fig. 4a). In the centromeres and TEs of *O. sativa*, Unimeth also profiles almost all the cytosines, while BS-seq only profiled 57.6% and 81.0% cytosines, respectively.

We further compared the site-level methylation frequencies obtained from nanopore-based methods with those from BS-seq in *O. sativa*. As shown in Fig. 4b, Unimeth consistently outperforms Dorado and DeepPlant in 5mC detection across all three contents in centromeres and TEs. Specifically, for CpG methylation detection, all three nanopore-based methods show comparable performance in TEs. In centromeres, Unimeth achieves the highest performance with a Pearson correlation of 0.9548 and a root mean square error (RMSE) of 0.0673. For CHG methylation detection, the Pearson correlations of Unimeth are more than 5% and 7% higher than those of the other two methods in TEs and centromeres, respectively. For CHH methylation detection, the advantage of Unimeth is even more pronounced, particularly in centromeres, where its Pearson correlation is 11% and 37% higher than those of DeepPlant and Dorado, respectively. We also evaluated the performance of Unimeth in 5mC detection in repetitive genomic regions of *A. thaliana* by using in-house generated data and the Col-CEN[13] reference genome (Supplementary Fig. 6). Unimeth alsodemonstrates the overall highest correlations and lowest RMSEs, highlighting its capability for accurate 5mC detection in repetitive genomic regions.

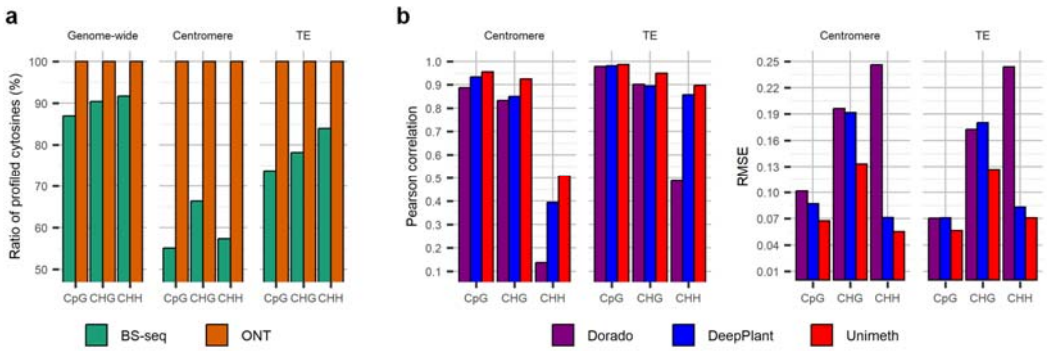

Fig. 4 5mC profiling in repetitive genomic regions in chromosome 1 of *O. sativa*. Analysis was performed using 73.7× WGBS reads and 99.2× ONT reads. **a** Ratios of profiled cytosines by BS-seq (Bismark) versus ONT (Unimeth). Only cytosines covered by at least 5 reads are counted. **b** Comparison of site-level methylation frequencies generated by Unimeth, Dorado, and DeepPlant with those from BS-seq, assessed by Pearson correlation and RMSE.

Unimeth improves 5mCpG detection on mammal genomes

Although DNA methylation in mammals occurs predominantly at CpG sites, existing methods trained on human genomes demonstrate limited generalizability to plant species, as evidenced in our earlier analyses. To establish a comprehensive benchmark, we evaluated their performance on human CpG detection tasks with Unimeth, a specialized variant trained exclusively on human data for CpG methylation detection.

For the R10.4.1 5kHz chemistry, we evaluated Unimeth against Dorado, Rockfish, and DeepMod2 using four human reference samples (HG001-HG004). As shown in Fig. 5a, all methods demonstrated high read-level performance measured by AUC, which aligns with expectations for within-species validation. Unimeth achieved the highest AUC of 0.9974, marginally outperforming Rockfish at 0.9956. For the site-level evaluation, as shown in Fig. 5b, Unimeth consistently showed superior Pearson correlations across all sequencing depths for all four datasets. While Dorado exhibited higher correlations at full depth for HG001 and HG003, Unimeth maintained better performance at lower depths, demonstrating its robustness across varying sequencing coverage.

Notably, Dorado's read-level performance was comparatively weaker, with an AUC of only 0.9759, suggesting limitations in distinguishing positive and negative samples.

To further test the performances of Unimeth, we further evaluated Unimeth using the nanopore data of mouse C57BL/6 cell line[26] and *E. coli*[34].For the C57BL/6 cell line, Unimeth maintained the highest accuracy, achieving a 0.9766 AUC on read-level and 0.8462 Pearson correlation on site-level (Supplementary Fig. 7).

Consistent with our findings in 5mC detection, Unimeth also achieved a substantial reduction in false positives for CpG site calling (Supplementary Fig. 8). Across the four human datasets and the mouse dataset, Unimeth lowered the minimum FPR to 0.99%, compared to the 1.21% achieved by Rockfish. To provide more definitive evidence, we further evaluated the models on an *E. coli* dataset treated with the methylation-sensitive restriction enzyme MSssl, which enzymatically removes all CpG methylation. In this controlled setting where any reported methylation constitutes a false positive, Rockfish yielded an FDR of 0.12%. Unimeth dramatically reduced this value to 0.004%, substantially outperforming all other methods and confirming its superior specificity.

We also trained specialized models for both R10.4.1 (4kHz) and R9.4.1 chemistries for comprehensive CpG detection. On R10.4.1 (4kHz), where Rockfish is unavailable, Unimeth was benchmarked against DeepMod2 and Dorado. It demonstrated superior performance across three datasets, achieving an average read-level AUC of 0.9977 and a site-level Pearson correlation of 0.9535. Notably, the comparison between 4kHz and 5kHz chemistries revealed comparable performance for Unimeth at both read and site levels, suggesting that nanopore sequencing frequency has minimal impact on methylation detection accuracy. For the older R9.4.1 chemistry, Unimeth also maintained strong performance, attaining a read-level AUC of 0.9953 and a site-level Pearson

correlation of 0.9411, making it the top-performing tool despite the technological constraints. However, its reduced performance relative to both R10.4.1 platforms (4kHz and 5kHz) at both analysis levels indicates that the improved signal sensitivity of R10.4.1 chemistry enables more accurate computational detection of methylation patterns.

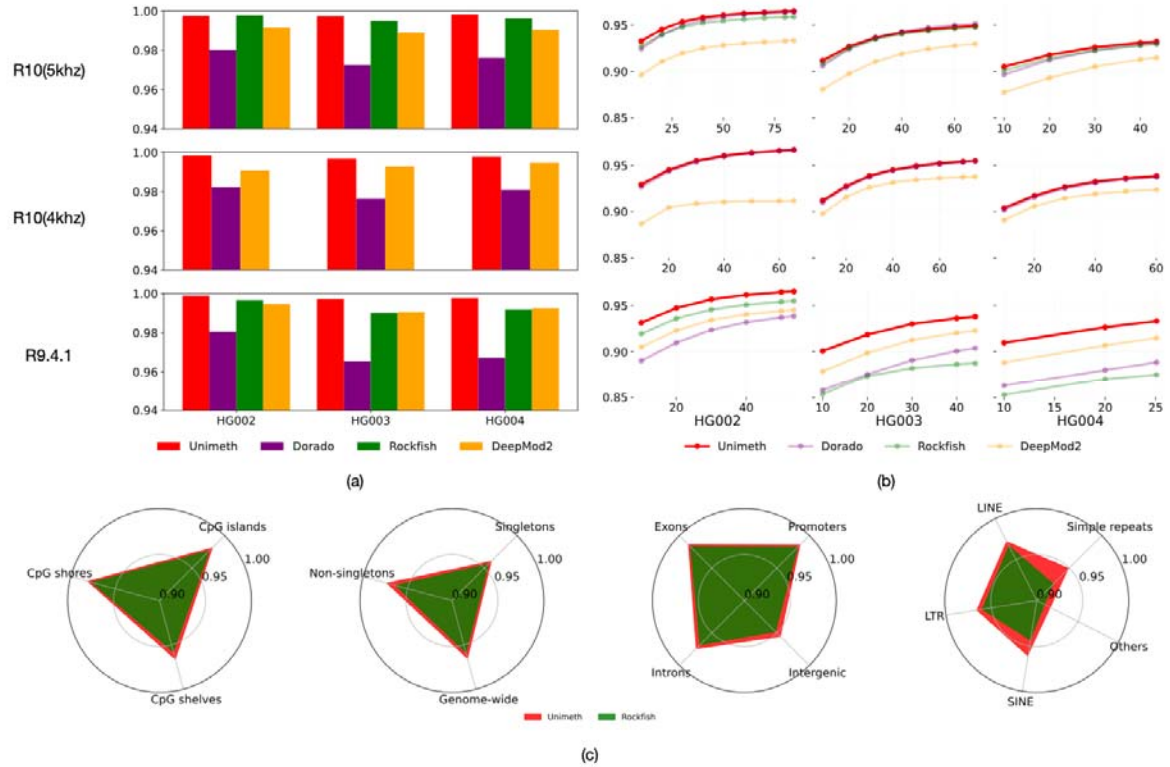

Fig. 5 Evaluation of CpG detection performance on human samples. **a** Read-level performance comparison of Unimeth with Dorado, Rockfish, and DeepMod2 across different sequencing chemistries (R10.4.1(5kHz), R10.4.1(4kHz), R9.4.1) and three human sample datasets (HG002, HG003, HG004). **b** Site-level performance of the three chemistries on the three datasets, showing the Pearson correlation with BS-seq results at varying sequencing depths. **c** Read-level and site-level performance in different genomic contexts and regions, categorized as CpG island regions, Singleton/Non-singleton, Genomic regions, and Repeats. Chromosome 1 of HG002 is used for evaluation, comparing Unimeth against Rockfish.

DNA methylation is strongly influenced by local CpG density, which correlates with distinct genomic functions and regulatory mechanisms. To systematically assess this relationship, we evaluated Unimeth against the second-best method, Rockfish, across different CpG density contexts using the HG002 dataset (Fig. 5c). We first assessed model performance across different CpG genomic resorts, including CpG islands (high-density regions), CpG shores (adjacent regions), and

CpG shelves (more distant regions). Unimeth demonstrated superior performance across all three regions, with particularly notable advantages in CpG islands and shores where methylation sites are more densely clustered. Beyond regional classification, CpG sites can also be categorized as singletons (isolated CpGs without close neighbors within a short genomic range) or non-singletons (clustered CpGs in close proximity). This distinction is biologically meaningful, as non-singleton sites often exhibit cooperative methylation patterns and are enriched in functional genomic regions. Our results demonstrate that Unimeth achieves superior performance on both singleton and non-singleton sites, with a more pronounced advantage on non-singletons. This enhanced capability likely stems from our model's capacity to process multiple methylation sites simultaneously within a genomic patch, enabling effective perception of CpG density and inter-site dependencies during prediction.

Beyond CpG density, methylation plays distinct functional roles in various genomic contexts. We therefore expanded our analysis to include functionally annotated regions, evaluating performance in promoters, exons, introns, and intergenic regions, where Unimeth demonstrated robust performance across all categories. Additionally, an important application of long-read sequencing lies in characterizing repetitive genomic regions. We examined various repeat element classes, including short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), long terminal repeats (LTRs), simple repeats, and other repetitive sequences. Across all these diverse functional and repetitive contexts, Unimeth consistently maintained its superior performance as the most accurate detection tool available. Notably, Unimeth achieved an average performance improvement of 1.2% in these complex repetitive regions and 2.37% in simple repeat regions, significantly higher than its overall performance improvement, demonstrating its particular strength

in resolving methylation patterns in genomic contexts where traditional methods often struggle.

Unimeth accurately detects 6mA and helps find more nucleosomes

While 5mC represents the predominant form of DNA methylation in eukaryotes, 6mA is the most prevalent methylation form in prokaryotes and contributes to biological processes such as genome repair[3,35]. In human genome, 6mA occurs at relatively low abundance, with an estimated abundance of approximately 0.05%[36]. Recent advances in Fiber-seq technology have enabled genome-wide profiling of open chromatin and nucleosome positioning at single-molecule resolution by leveraging m6A methyltransferase to label accessible regions[31,37]. This approach provides detailed structural insights into chromatin organization. Here, using the ONT Fiber-seq data of HG002 from Anupama *et al*.[31], we trained and evaluated Unimeth in DNA 6mA detection (Methods). Additionally, we tested the 6mA model of Unimeth on the ONT data of *E. coli* obtained from Kulkarni *et al*.[34]

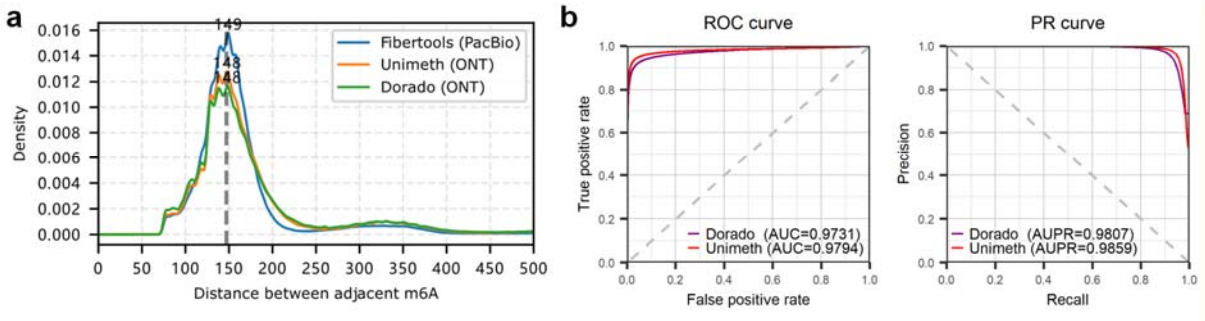

Fig. 6 Evaluation of Unimeth on DNA 6mA detection. **a** Density of the distance between adjacent m6A on the same chromatin fiber for the ONT and PacBio Fiber-seq data of HG002. **b** ROC and PR curves of the read-level 6mA detection on the *E. coli* dataset in GATC contexts.

For the Fiber-seq data of HG002, we used the nucleosome identified by PacBio sequencing as the gold standard, processed through the fibertools pipeline[31]. Unimeth identified 6.87% of sites as methylated, lower than the 7.63% detected by Dorado. Despite calling fewer methylated sites, Unimeth led to the identification of more nucleosomes. In a sampled set of 10,000 reads, our method detected 132,155 nucleosomes compared to 131,362 by Dorado, suggesting higher precision in 6mA

detection. As shown in Fig. 6a, the nucleosome length distributions derived from fibertools (PacBio), Dorado, and Unimeth all center near the theoretical value of 147 bp. While both nanopore-based methods produce peaks around 148 bp, Unimeth exhibits a distinctly sharper and higher density at this length, resulting in a distribution that aligns more closely with the result of PacBio sequencing. This refined distribution profile not only reflects more consistent fragment sizing but also corroborates the improved accuracy of our method in nucleosome detection.

The *E. coli* dataset from Kulkarni *et al*.[34] contains ONT reads of two strains: one is the wild-type strain (WT), in which nearly 100% adenines in GATC motif are methylated, and the other is a mutant strain (DM), in which the methylases *dam* were knocked out to deplete 6mAs. At the read level across the 3.9 million tested sites, Unimeth achieved an AUC of 0.9794 and an AUPR of 0.9859, slightly outperforming Dorado, which yielded an AUC of 0.9731 and an AUPR of 0.9807 (Fig. 6b).

Discussion

In this study, we introduced Unimeth, a unified deep learning framework for accurate and efficient detection of DNA methylation from nanopore sequencing data. Departing from the conventional k-mer-based prediction paradigm, Unimeth employs a novel patch-based transformer architecture that captures contextual dependencies among neighboring methylation sites, enabling simultaneous prediction of multiple sites within a genomic segment. This design not only improves predictive accuracy but also significantly enhances computational efficiency. Furthermore, Unimeth adopts a site-level training strategy, which enables robust predictions not only for fully methylated and unmethylated sites, but also for partially methylated ones, leading to a comprehensive improvement in site-level methylation calling.

Unimeth achieves consistently high performance for methylation detection of multiple cytosine contexts (CpG, CHG, CHH), in both plant and mammalian genomes, accommodates various sequencing chemistries (R9.4.1, R10.4.1 4kHz and 5kHz), and generalizes robustly to unseen species and batch variations. A particular challenge in plant epigenomics is the detection of rare methylation events in CHH contexts, where strong class imbalance has traditionally hindered machine learning performance. Previous methods often relied on dataset rebalancing to mitigate this issue, but such approaches distort the natural methylation distribution and tend to introduce substantial false positives during inference. In contrast, Unimeth is trained on unaltered data distributions and demonstrates markedly reduced false positive rates in CHH contexts. This improved specificity is validated not only on native sequencing data, but also on methylation-depleted controls (e.g., using ddcc and met1-3 enzyme systems), confirming the model's high reliability in real-world applications.

Additionally, our repeat region analyses reveal that bisulfite-based methods suffer from serious alignment artifacts in repetitive genomic regions, limiting their accuracy in these important contexts. Leveraging nanopore sequencing and Unimeth, we successfully achieve precise methylation detection in such regions. Beyond 5mC detection, Unimeth also outperforms existing methods in 6mA identification. When applied to Fiber-seq data, it accurately identifies a higher number of nucleosome positions with improved distribution consistency relative to experimental standards.

The framework's ability to perform consistently across technologies and species—without requiring dataset rebalancing or extensive preprocessing—underscores its practical utility and biological relevance. As nanopore sequencing continues to evolve, Unimeth provides a flexible, accurate, and scalable solution for DNA methylation analysis, paving the way for more integrated and comprehensive epigenomic profiling in the future.

Methods

Preparation of plant materials and DNA extraction

Wild-type Arabidopsis thaliana (L.) Heynh. Columbia-0 (Col-0), Oryza sativa L. ssp. Japonica cv. Nipponbare, and *Musa* spp. 'Y137' were used in this study. Seeds of *A. thaliana* were surface sterilized in 4% sodium hypochlorite, vernalized for 2 days at 4 °C, and grown on half Murashige & Skoog (MS) plates for 21 days at 22 °C, 70% relative humidity with a 16 h/8 h light/dark regime. Seeds of *O. sativa* were sown in soil and cultured for 28 days in a growth chamber at 28 °C, 70% relative humidity with a 16 h/8 h light/dark regime. The harvested leaves of both *A. thaliana* and *O. sativa* were quickly frozen with liquid nitrogen, stored at -80 °C until further use. The materials of *Musa* spp. 'Y137' were provided by Wuhan Frasergen Bioinformatics Co., Ltd. The genomic DNA of each sample was extracted by QIAGEN® Genomic DNA extraction kit (Cat#13323) according to the manufacturer's standard operating procedure. The extracted DNA was detected by NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific) for DNA purity.

Nanopore sequencing and BS-seq

We generated nanopore and BS-seq reads of *A. thaliana* Col-0, *O. sativa* Japonica, and *Musa* spp. 'Y137' in this study. For nanopore sequencing, the extracted genomic DNA was qualified, size-selected using the BluePippin system (Sage Science), and then end-repaired. PCR adapters supplied in the nanopore sequencing kit (SQK-LSK114) were ligated to the end-repaired DNA. Finally, Qubit® 4.0 Fluorometer (Invitrogen) was used to quantify the size of library fragments. To generate nanopore reads, the prepared libraries are loaded into flow cells (R10.4.1 5kHz, FLO-PRO114M) of a PromethION sequencer (ONT). For BS-seq, the extracted genomic DNA was

first sheared by Covaris and purified to 200–350 bp (base pair) in average size. Then the sheared DNA was end-repaired and ligated to methylated sequencing adapters. The adapter-ligated DNA was bisulfite-converted and PCR-amplified. The bisulfite conversion kit and library preparation kit were TIANGEN DNA Bisulfite Conversion Kit (cat #: DP215, TIANGEN BIOTECH) and TruSeq DNA Methylation Kit (cat #: EGMK91324, Illumina), respectively. The libraries of the samples were sequenced on a NovaSeq6000 sequencer (Illumina), finally to obtain pair-end 150 bp reads.

Public nanopore, BS-seq, and EM-seq data

We downloaded nanopore reads of 12 species, including 9 plants (*S. miltiorrhiza*, *S. tuberosum*, *R. communis*, *V. vinifera*, *B. vulgaris*, *C. sinensis*, *S. lycopersicum*, *A. thaliana*, and *O. sativa*), 2 mammals (*H. sapiens* and *M. musculus*), and 1 bacterium (*E. coli*), from publicly available datasets. The downloaded datasets include nanopore reads sequenced using R10.4.1 5kHz, R10.4.1 4kHz, and R9.4.1 kits. For the R10.4.1 5kHz kit, we got nanopore reads of 9 plant species from Chen *et al.*[27] and got nanopore reads of 4 *H. sapiens* samples (HG001-HG004) from the ONT Open Data project (https://epi2me.nanoporetech.com/dataindex/). We also got nanopore reads of *M. musculus* C57BL/6 from Stanojević *et al.*[26] and got nanopore reads of *E. coli* generated by Kulkarni *et al.*[34]. The data of *E. coli* contains reads of wild-type strain (WT), *dam*-*dcm*-methylase-knocked-out strain (DM), and M.SssI-treated strain (MSssI). For the R10.4.1 4kHz kit, we got nanopore reads of 3 *H. sapiens* samples (HG002-HG004) from the ONT Open Data project. For the R9.4.1 kit, we got nanopore reads of HG002-HG004 from the ONT Open Data project, and got reads of HG001 from Jain *et al.*[38] (See Supplementary Data 1). We also got reads of *A. thaliana* and *O. sativa* from our previous study[30].

To assess Unimeth, we also downloaded the corresponding BS-seq (or EM-seq) data of the used

species/samples. For *H. sapiens*, we downloaded the BS-seq data of HG002 from the ONT Open Data project and the EM-seq data of HG001, HG003, and HG004 from the SEQC2 study[11]. The BS-seq data of *M. musculus* were downloaded from Stanojević *et al.*[26]. For the 9 plant species, we got the corresponding BS-seq data from Chen *et al.*[27] and Ni *et al.*[30].

Reference genomes and annotations

For *A. thaliana* and *O.sativa*, we used Col-CEN[13] and T2T-NIP[33] (AGIS1.0) as the reference genomes, respectively. We also downloaded the corresponding gene and centromere annotations of these two genomes. For *Musa* spp., we generated an assembly as the reference with hifiasm[39] (v0.25.0-r726) and purge_dups[40] (v1.2.6). For the other 7 used plant samples, we got the same reference genomes as Chen *et al.*[27]. To process the *H. sapiens* data, we used CHM13 v2.0 as the reference genome[41]. The gene annotation was downloaded from the GitHub repository marbl/CHM13 [https://github.com/marbl/CHM13]. The annotations of CpG islands, peri/centromeric satellites, and repetitive genomic elements (RepeatMasker) were downloaded from corresponding tracks of UCSC Genome Browser (T2T CHM13v2.0/hs1). For *M. musculus*, we downloaded the GRCm39 (Release M37) genome from GENCODE[42]. The reference genome of *E. coli* was downloaded from NCBI (GCF_000005845.2). Details of the reference genomes are provided in Supplementary Note 1.

Processing of BS-seq and EM-seq data

We applied the Nextflow pipeline nf-core/methylseq[43] (2.5.0) to process all the BS-seq and EM-seq data. We ran the Bismark[44] workflow of methylseq with parameters "--comprehensive true --cytosine_report true" to get the strand-specific methylation information of all three motifs (CpG, CHG, and CHH). The Biamark workflow outputs per-read methylation calls for each cytosine. The

methylation frequencies of cytosines are then calculated and saved in bedmethyl format. For the samples with multiple replicates (*H. sapiens* HG001, HG003, and HG004, *M. musculus* C57BL/6, and *A. thaliana* Col-0, see Supplementary Data 2), we calculated the average methylation frequencies of all replicates as the final results. We only kept the sites with at least 5 reads covered in all replicates for further evaluation. Details of the used BS-seq and EM-seq data, including the read depth and the methylation levels, are listed in Supplementary Data 2.

Processing of nanopore data

The nanopore data used in this study includes R10.4.1 5kHz, R10.4.1 4kHz, and R9.4.1 reads in POD5 or FAST5 files. For 5mC detection, we used Dorado[29] (v0.7.1), the official basecaller of ONT, to basecall all the nanopore raw reads. To keep the signal-to-sequence alignment information, we added the parameter "--emit-moves" in the Dorado basecalling command. The raw reads in FAST5 format were converted to POD5 format before basecalling for optimal basecalling performance. We used the baseacalling models "dna_r10.4.1_e8.2_400bps_sup@v5.0.0", "dna_r10.4.1_e8.2_400bps_sup@v4.1.0", and "dna_r9.4.1_e8_sup@v3.3" for the R10.4.1 5kHz, R10.4.1 4kHz, and R9.4.1 reads, respectively. For 6mA detection from the nanopore reads of *E. coli* WT and DM strain, and the HG002 fiber-seq data, we used Dorado (v0.9.1) with the model "dna_r10.4.1_e8.2_400bps_sup@v5.0.0_6mA@v3" for bsaecalling. The basecalling of Dorado integrates minimap2[45] to perform the read-to-reference alignment simultaneously. The results of basecalling were saved in BAM format. Details of the used nanopore data, including the sequencing kit, number of reads, read length, read depth, and read quality, are provided in Supplementary Data 3. We split the nanopore reads based on the chromosomes/contigs for training and evaluation. Details of the data partition are provided in Supplementary Data 4.

To evaluate Unimeth, we performed methylation calling with the nanopore data using Dorado (v0.7.1), Rockfish (v0.3), DeepMod2 (v0.3.1), and DeepPlant (v1.1.0) for comparison. The methylation calling of Dorado was performed during the basecalling step. To call methylation from R10.4.1 5kHz, R10.4.1 4kHz, and R9.4.1 reads with Dorado, we used the modifications models "dna_r10.4.1_e8.2_400bps_sup@v5.0.0_5mC_5hmC@v1", "dna_r10.4.1_e8.2_400bps_sup@v4.1.0_5mCG_5hmCG@v2", and "dna_r9.4.1_e8_sup@v3.3_5mCG@v0.1", respectively. We ran Rockfish with models "rf-r10.4.1-5kHz-base.pt" and "rf-base.pt" to call CpG methylation from R10.4.1 5kHz and R9.4.1 reads. Note that before calling methylation from the R9.4.1 reads using rockfish, we basecalled the reads using the Guppy basecaller (v6.2.1) with the model config "dna_r9.4.1_450bps_sup_prom.cfg". For DeepMod2, we used the models "bilstm_r10.4.1_5khz_v5.0", "bilstm_r10.4.1_4khz_v4.1", and "bilstm_r9.4.1" to call CpG methylation from R10.4.1 5kHz, R10.4.1 4kHz, and R9.4.1 reads, respectively. We used DeepPlant with the default R10.4.1 5kHz model and parameters to call 5mC methylation from the nanopore reads of the plant samples. As Chen *et al.* did, before using DeepPlant, we used Dorado with the model "dna_r10.4.1_e8.2_400bps_hac@v5.0.0" for basecalling. We used DeepSignal-plant (v0.1.6) and the recommended pipeline (Guppy (v6.4.2) for basecalling, Tombo (v1.5.1) for re-squiggle) to call 5mC methylation from the R9.4.1 reads of *A. thaliana* and *O. sativa* for comparison.

Unimeth model architecture

Unimeth takes raw nanopore signals and their corresponding basecalled genomic sequences as input. For R10.4.1 reads, we got the signal-to-sequence alignments based on the move tables of the BAM files. For R9.4.1 reads, we re-squiggled the signals to reference genomes (Supplementary Note 2).

Basecalled reads are divided into fixed-length subsequences of 256 bp, and the signal segment aligned to each subsequence is extracted using a mapping table which we refer to these paired signal-sequence segments as "patches". Within each patch, special methylation tokens ([CpG], [CHG], [CHH]) are inserted into the DNA sequence following cytosine bases located in corresponding sequence contexts.

The model architecture consists of an encoder and a decoder. The encoder is designed to learn representations from raw signal inputs and comprises three core components: a signal processor, a positional encoder, and a transformer backbone. The signal processor normalizes the raw signals and uses two CNN layers to convert them into signal embeddings. The positional encoder incorporates base position information using sinusoidal positional encoding, which can be formulated as:

$$PE(pos,\ 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right)$$

$$PE(pos,\ 2i+1) = \cos\left(\frac{pos}{10000^{2i/d}}\right)$$

where *pos* denotes the signal position, *i* is the dimension index, and *d* is the embedding dimension. These signal and positional embeddings are combined and passed through 12 layers of transformer encoder to produce the final signal embeddings.

The decoder processes the tokenized DNA sequence with inserted methylation markers. It consists of a token embedding layer, a positional encoder, and a transformer backbone. The tokenizer converts the sequence and special tokens into token embeddings. The positional encoder, using the same sinusoidal encoding scheme as above, injects positional information into these embeddings. The decoder backbone contains 12 layers of transformer decoder that attend to both the token embeddings and the signal embeddings from the encoder.

The output of the decoder maintains the original patch length and structure: it reconstructs the

nucleotide bases and replaces each methylation token with a prediction token, either '[+]' indicating a methylated site, or '[-]' indicating an unmethylated site.

Training process of Unimeth

Unimeth was trained in three phases: pre-training, fine-tuning, and calibration. During pre-training, the model learns basecalling by taking raw nanopore signals as input and generating the corresponding nucleotide sequences without inserting special tokens. This phase employs a next-token prediction objective, where the loss is formulated using standard cross-entropy:

$$L_{pre-train} = -\frac{1}{N}\sum_{i=1}^{N} \log P\left(x_i \mid x_1, \ldots, x_{i-1}; \theta\right)$$

This process enables the model to capture fundamental relationships between signal patterns and nucleotide identities, providing a strong initialization for efficient convergence in subsequent methylation prediction tasks.

In the fine-tuning phase, the model is adapted for methylation prediction. The key modification from pre-training is the introduction of special tokens into the target labels. For example, at a CpG site, the original sequence “C G” is transformed into “C [CpG] G”, where the token “[CpG]” marks the methylation context and the position to be predicted. During this stage, the causal mask is removed, and the full sequence serves as input to the decoder. The model is trained to predict either “[+]” (methylated) or “[–]” (unmethylated) at each marked site, producing outputs such as “C [+] G” or “C [–] G”.

To accommodate the continuous nature of BS-seq sequencing methylation ratios (ranging from 0 to 1), we relax the conventional binary thresholding. Sites with methylation levels below 20% are treated as negative, and those above 80% as positive. We introduce a weighted loss function that

assigns higher importance to samples closer to the extremes (0 or 100). The weight for predicting base is fixed as 1 and for methylation is defined as:

$$w = \left(\frac{|b - 0.5|}{0.5}\right)^2$$

where b is the BS-seq methylation ratio. The loss for each methylation type is computed as the weighted cross-entropy between the prediction and the binarized label. For plant models, which involve three methylation types (CpG, CHG, CHH), the total loss is the sum of the weighted losses for each type. For the human model, which only predicts CpG methylation, standard weighted cross-entropy is applied.

To leverage the model's ranking capability and the site-level methylation ratio from bisulfite sequencing, a calibration procedure was performed to generate refined, site-aware labels for the training set.

Inference and grouping. The fine-tuned model was applied to all reads in the training set to obtain a methylation probability $p_{\text{methyl}} \in [0,1]$ for each read at each cytosine site. Reads were then grouped by their genomic site s.

Site-level reference calculation. For each site s, the bisulfite-seq-derived methylation ratio ρ_s and the total number of overlapping reads N_s were used to calculate the expected number of methylated reads:

$$n_s^+ = \lfloor \rho_s \cdot N_s \rfloor$$

The expected number of unmethylated reads is

$$n_s^- = N_s - n_s^+$$

Rank-based label assignment. For the set of reads r_i at site s, all reads were sorted in descending order of their predicted probability $p_{\text{methyl}}(r_i)$. The top n_s^+ reads were assigned a

calibrated label $\widetilde{y_i} = 1$ (methylated). The remaining n_s^- reads were assigned $\widetilde{y_i} = 1$ (unmethylated). This ensures the proportion of reads labeled as methylated at the site matches the bulk bisulfite-seq measurement ρ_s.

Formally, for sorted reads where $p_{\text{methyl}}(r_1) \geq p_{\text{methyl}}(r_2) \geq \cdots \geq p_{\text{methyl}}(r_{N_s})$:

$$\widetilde{y_i} = \begin{cases} 1, & i \leq n_s^+ \\ 0, & i > n_s^+ \end{cases}$$

Re-training: The model was then re-trained (or further fine-tuned) using the original input features (currents, squiggles) but with the newly assigned, calibrated labels $\widetilde{y_i}$ for reads from one or two held-out chromosomes. This step teaches the model to align its read-level predictions more closely with the site-level biological truth, thereby improving its accuracy on partially methylated sites.

This calibration method is based on the premise that while the absolute probability scores from the initial model may not be perfectly thresholded, their relative ranking within a site is reliable. The procedure effectively uses the high-confidence site-level ratio ρ_s as a constraint to "correct" noisy read-level labels and include information from partially methylated sites.

During the pre-training phase, three separate models were trained for the R9.4.1, R10.4.1 (4 kHz), and R10.4.1 (5 kHz) chemistries using human genome data. In the fine-tuning and calibration phases, task-specific models were developed for different detection targets. For 5mC detection in plants, the R10.4.1 model was trained on *S. miltiorrhiza*, *S. tuberosum*, and *R. communis*, in accordance with the training dataset used in DeepPlant[27]. The R9.4.1 model was trained on *A. thaliana* and *O. sativa*, following the experimental setup of DeepSignal-Plant[30]. For CpG detection in human data, three chemistry-specific models were fine-tuned using the HG002 sample with each model trained on R9.4.1, R10.4.1 (4 kHz), and R10.4.1 (5 kHz) data, respectively. To prevent data leakage, chromosome 1 was excluded from the training set in all experiments, ensuring that test regions

remained unseen during model training.

In our experiment, we use a learning rate of 1e-4 for pretraining, and use 1e-5 at finetuning and calibration phase. Running average coefficients were set at their default values $\beta_1 = 0.9$, $\beta_2 = 0.999$.

Evaluation

Evaluation at read level. For the human and plant samples, we first select high-confidence methylated and unmethylated sites from the results of BS-seq. Then we extract positive and negative samples of the selected sites from CCS reads. We calculate accuracy, sensitivity, specificity, Area Under the Curve (AUC), and AUPR based on the prediction of all tested sites. For the *E. coli* dataset, we extract positive and negative samples from the reads of wild-type and M.SssI-treated DNA, respectively.

Evaluation at site level. We evaluate Unimeth at the genome-wide site level by comparing per-site methylation frequencies predicted by Unimeth with the results of BS-seq. Using the methylation frequencies of CpGs, Pearson correlation (r), the coefficient of determination (r^2), Spearman correlation (ρ), and root mean square error (RMSE) are calculated. For each comparison, we only compare CpGs that have at least 5× coverage in the results of both nanopore sequencing and BS-seq. To evaluate the methylation frequencies predicted by Unimeth under different coverage of nanopore reads, we randomly subsample the nanopore reads using rasusa[46] (v2.0.0).

Data availability

All sequencing data generated in this study, including the ONT R10.4.1 5kHz and BS-seq reads of *A. thaliana*, *O. sativa*, and *Musa* spp. will be uploaded to the GSA (Genome Sequence Archive) of

NGDC (National Genomics Data Center).

The ONT R10.4.1 5kHz nanopore reads of HG001, HG002, HG003, and HG004 are available at the ONT Open Data project [https://epi2me.nanoporetech.com/giab-2023.05/]. The ONT R10.4.1 4kHz nanopore reads of HG002, HG003, and HG004 are available at the ONT Open Data project [https://epi2me.nanoporetech.com/askenazi-kit14-2022-12/]. The R9.4.1 nanopore reads of HG002 are available at the ONT Open Data project [https://labs.epi2me.io/gm24385_2020.11/] with the flowcell ID PAG07165. The BS-seq reads of HG002 are available at the ONT Open Data project [https://labs.epi2me.io/gm24385-5mc/]. The ONT Fiber-seq reads of HG002 are available at the NCBI BioProject database under accession PRJNA956114 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA956114].

The ONT R10.4.1 5kHz nanopore reads of *S. miltiorrhiza*, *S. tuberosum*, *R. communis*, *C. sinensis*, *S. lycopersicum*, *V. vinifera*, *A. thaliana*, *O. sativa*, and *B. vulgaris* generated by Chen *et al*.[27], together with the corresponding BS-seq reads, are available in the GSA of NGDC under accessions PRJCA030666 [https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA030666] and PRJCA023349 [https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA023349]. The ONT R9.4.1 reads of *A. thaliana* and *O. sativa*, as well as the corresponding BS-seq reads, are available in the GSA of NGDC under accession PRJCA004326 [https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA004326].

The ONT R10.4.1 5kHz nanopore reads of *E. coli* from Kulkarni *et al*.[34] are available from AWS [https://registry.opendata.aws/ont_basemod_data/].

Code availability

Unimeth is publicly available at GitHub [https://github.com/sekeyWang/unimeth].

References

1 Mattei, A. L., Bailly, N. & Meissner, A. DNA methylation: a historical perspective. *Trends in Genetics* **38**, 676-707 (2022).

2 Boulias, K. & Greer, E. L. Means, mechanisms and consequences of adenine methylation in DNA. *Nature Reviews Genetics* **23**, 411-428, doi:10.1038/s41576-022-00456-x (2022).

3 Beaulaurier, J., Schadt, E. E. & Fang, G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nature Reviews Genetics* **20**, 157-172, doi:10.1038/s41576-018-0081-3 (2019).

4 Kong, Y., Mead, E. A. & Fang, G. Navigating the pitfalls of mapping DNA and RNA modifications. *Nature Reviews Genetics* **24**, 363-381, doi:10.1038/s41576-022-00559-5 (2023).

5 Liu, T. & Conesa, A. Profiling the epigenome using long-read sequencing. *Nature Genetics*, doi:10.1038/s41588-024-02038-5 (2025).

6 Taiwo, O. *et al.* Methylome analysis using MeDIP-seq with low DNA concentrations. *Nature Protocols* **7**, 617-636, doi:10.1038/nprot.2012.012 (2012).

7 Krueger, F., Kreck, B., Franke, A. & Andrews, S. R. DNA methylome analysis using short bisulfite sequencing data. *Nature Methods* **9**, 145-151, doi:10.1038/nmeth.1828 (2012).

8 Vaisvila, R. *et al.* Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Research* **31**, 1280-1289, doi:10.1101/gr.266551.120 (2021).

9 Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* **21**, 30, doi:10.1186/s13059-020-1935-5 (2020).

10 Fu, Y., Timp, W. & Sedlazeck, F. J. Computational analysis of DNA methylation from long-read sequencing. *Nature Reviews Genetics*, doi:10.1038/s41576-025-00822-5 (2025).

11 Foox, J. *et al.* The SEQC2 epigenomics quality control (EpiQC) study. *Genome Biology* **22**, 332, doi:10.1186/s13059-021-02529-2 (2021).

12 Baduel, P. *et al.* Transposable elements are vectors of recurrent transgenerational epigenetic inheritance. *Science* **0**, eady3475, doi:10.1126/science.ady3475 (2025).

13 Naish, M. *et al.* The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science* **374**, eabi7489, doi:10.1126/science.abi7489 (2021).

14 Chen, W. *et al.* Pan-centromere landscape and dynamic evolution in Brassica plants. *Nature Plants*, doi:10.1038/s41477-025-02131-5 (2025).

15 Ni, P. *et al.* DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. *Nature Communications* **14**, 4054, doi:10.1038/s41467-023-39784-9 (2023).

16 Akbari, V. *et al.* Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase. *Genome Biology* **22**, 68, doi:10.1186/s13059-021-02283-5 (2021).

17 Kong, Y. *et al.* Critical assessment of nanopore sequencing for the detection of multiple forms of DNA modifications. *bioRxiv*, 2024.2011.2019.624260, doi:10.1101/2024.11.19.624260 (2024).

18 Yuen, Z. W.-S. *et al.* Systematic benchmarking of tools for CpG methylation detection from Nanopore sequencing. *Nature Communications* **12**, 1-12 (2021).

19 Liu, Y. *et al.* DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biology* **22**, 295 (2021).

20 Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology* **39**, 1348-1365, doi:10.1038/s41587-021-01108-x (2021).

21 Stoiber, M. *et al.* De novo Identification of DNA Modifications Enabled by Genome-Guided

Nanopore Signal Processing. *bioRxiv*, 094672 (2017).

22 Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods* **14**, 407-410 (2017).

23 Ni, P. *et al.* DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics* **35**, 4586-4595, doi:10.1093/bioinformatics/btz276 (2019).

24 Technologies, O. N. *Taiyaki*, <https://github.com/nanoporetech/taiyaki> (2019).

25 Ahsan, M. U., Gouru, A., Chan, J., Zhou, W. & Wang, K. A signal processing and deep learning framework for methylation detection using Oxford Nanopore sequencing. *Nature Communications* **15**, 1448, doi:10.1038/s41467-024-45778-y (2024).

26 Stanojević, D., Li, Z., Bakić, S., Foo, R. & Šikić, M. Rockfish: A transformer-based model for accurate 5-methylcytosine prediction from nanopore sequencing. *Nature Communications* **15**, 5580, doi:10.1038/s41467-024-49847-0 (2024).

27 Chen, H.-X. *et al.* Accurate cross-species 5mC detection for Oxford Nanopore sequencing in plants with DeepPlant. *Nature Communications* **16**, 3227-3227 (2025).

28 McIntyre, A. B. R. *et al.* Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. *Nature Communications* **10**, 579, doi:10.1038/s41467-019-08289-9 (2019).

29 Technologies, O. N. *Dorado*, <https://github.com/nanoporetech/dorado> (2022).

30 Ni, P. *et al.* Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning. *Nature Communications* **12**, 5976, doi:10.1038/s41467-021-26278-9 (2021).

31 Jha, A. *et al.* DNA-m6A calling and integrated long-read epigenetic and genetic analysis with fibertools. *Genome Research* **34**, 1976-1986 (2024).

32 Mo, W. *et al.* Single-molecule targeted accessibility and methylation sequencing of centromeres, telomeres and rDNAs in Arabidopsis. *Nature Plants* **9**, 1439-1450, doi:10.1038/s41477-023-01498-7 (2023).

33 Shang, L. *et al.* A complete assembly of the rice Nipponbare reference genome. *Molecular Plant* **16**, 1232-1236, doi:https://doi.org/10.1016/j.molp.2023.08.003 (2023).

34 Kulkarni, O. *et al.* Comprehensive benchmarking of tools for nanopore-based detection of DNA methylation. *bioRxiv*, 2024.2011.2009.622763, doi:10.1101/2024.11.09.622763 (2024).

35 Tourancheau, A., Mead, E. A., Zhang, X.-S. & Fang, G. Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nature Methods* **18**, 491-498, doi:10.1038/s41592-021-01109-3 (2021).

36 Xiao, C.-L. *et al.* N6-Methyladenine DNA Modification in the Human Genome. *Molecular Cell* **71**, 306-318.e307, doi:https://doi.org/10.1016/j.molcel.2018.06.015 (2018).

37 Stergachis, A. B., Debo, B. M., Haugen, E., Churchman, L. S. & Stamatoyannopoulos, J. A. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**, 1449-1454, doi:10.1126/science.aaz1646 (2020).

38 Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**, 338-345, doi:10.1038/nbt.4060 (2018).

39 Cheng, H. *et al.* Efficient near telomere-to-telomere assembly of Nanopore Simplex reads. *bioRxiv*, 2025.2004.2014.648685, doi:10.1101/2025.04.14.648685 (2025).

40 Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896-2898, doi:10.1093/bioinformatics/btaa025 (2020).

41 Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44-53, doi:10.1126/science.abj6987 (2022).

42 Mudge, Jonathan M. *et al.* GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Research* **53**, D966-D975, doi:10.1093/nar/gkae1078 (2025).

43 nf-core/methylseq: Crema Catalana (Zenodo, 2025).

44 Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572, doi:10.1093/bioinformatics/btr167 (2011).

45 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).

46 Hall, M. B. Rasusa: Randomly subsample sequencing reads to a specified coverage. *Journal of Open Source Software* **7**, 3941 (2022).

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants (Nos. 62350004, 62332020, 62302526) and the Project of Xiangjiang Laboratory (No. 23XJ01011). We thank Onkar Kulkarni and Divya Tej Sowpati from CSIR Centre for Cellular and Molecular Biology for providing the ONT data of *E. coli*. We thank Andrew B. Stergach and Mitchell R. Vollger from University of Washington for providing the PacBio and ONT Fiber-seq data of HG002. We thank Jonathan Cahn from Cold Spring Harbor Laboratory for providing their ONT data of *A. thaliana* mutants. We also thank Pierre Baduel from Université Paris Sciences et Lettres, Kateryna Makova from Penn State University, and Brandon D. Pickett from National Human Genome Research Institute (NHGRI) for providing their data and useful discussions. This work was carried out in part using computing resources at the High-Performance Computing Center of Central South University.

Author contributions

P.N. and S.K.W. conceived and designed this project. S.K.W., P.N., Y.F.X., and T.S. designed and implemented Unimeth, and evaluated Unimeth using the sequencing data. N.H. helped design the model of Unimeth. J.X.Z. and Y.S. generated the ONT data of the *A. thaliana* mutants and helped analyze the data. F.L. helped design the model and the experiments. All authors have read and approved the final version of this paper.

Competing interests

The authors declare no competing interests.