DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing

Peng Ni[1,2], Jinrui Xu[1,2], Zeyu Zhong[1,2], Jun Zhang[1,2], Neng Huang[1,2], Fan Nie[1,2], Feng Luo[3,*], Jianxin Wang[1,2,*]

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

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

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

[*]To whom correspondence should be addressed:

luofeng@clemson.edu, jxwang@mail.csu.edu.cn

Abstract. It has been reported recently that DNA 5-methylcytosine (5mC) in CpG contexts can be detected using PacBio circular consensus sequencing (CCS). However, the accuracy and robustness of computational methods using long CCS reads still need to be improved. In this study, we present a deep learning method, ccsmeth, to detect DNA 5mCpGs from PacBio CCS subreads. ccsmeth utilizes attention-based bidirectional Gated Recurrent Unit (GRU) networks to infer DNA methylation states. Testing ccsmeth using CCS subreads of amplified DNA and M.SssI-treated DNA, we found that ccsmeth achieved higher performances than existing methods. We also compared the results of ccsmeth on long CCS reads with bisulfite sequencing and Nanopore sequencing. The results demonstrated that ccsmeth can accurately detect 5mCpGs from CCS data sequenced using >10 kb insert library. Moreover, using PacBio CCS data, we proposed a pipeline which can detect haplotype-aware methylation in human.

Keywords: 5-methylcytosine, circular consensus sequencing, deep learning, haplotype-aware methylation

Introduction

5-methylcytosine (5mC), the most common form of DNA methylation, is involved in regulating many biological processes[1]. In humans, most 5mCs occur at CpG sites, which are associated with cancer[2], embryonic development[3] and aging[4]. Bisulfite sequencing (BS-seq) is now the most widely used methodology for profiling 5mC methylation[5,6]. In a bisulfite treated genomic DNA, unmethylated cytosines are converted to uracils, while methylated cytosines are unchanged[7]. Thus, methylation status of a segment of DNA can be yielded at single-nucleotide resolution. However, bisulfite treatment damages the DNA, which further leads to DNA degradation and the loss of sequencing diversity[8]. Recently, two bisulfite-free methods, ten-eleven translocation-assisted pyridine borane sequencing[9] (TAPS) and enzymatic methyl-seq[10] (EM-seq), were also developed, which are both reported to have more uniformly coverage and higher unique mapping rates than BS-seq. The same as BS-seq, TAPS and EM-seq can be applied to both short-read sequencing and long-read sequencing[11-14]. However, all these methods need extra laboratory techniques, which further leads to extra sequencing costs.

Two major long-read sequencing technologies, PacBio single-molecule real-time (SMRT) sequencing and Nanopore sequencing, can directly sequence native DNA without PCR amplification[15,16]. DNA base modifications alter polymerase kinetics in SMRT sequencing, and affect the electrical current signals near the modified bases in Nanopore sequencing[16]. Thus, DNA base modifications can be directly detected from

native DNA reads of SMRT and Nanopore sequencing without extra laboratory techniques[15,16]. For Nanopore sequencing, computational methods for 5mC detection either apply statistical tests to compare current signals of native DNA reads with an unmodified control (Tombo[17]), or use pre-trained Hidden Markov models (nanopolish[18]) and deep neural network models (Megalodon[19], DeepSignal[20]) without a control dataset. Previous studies have shown that methods using pre-trained models achieved high accuracies for DNA 5mC detection from human Nanopore reads[21,22].

Pulse signals in SMRT sequencing, which are associated with the nucleotides in which the polymerization reaction is occurring[16,23], includes the interpulse duration (IPD) and the pulse width (PW). IPD represents the time duration between two consecutive sequenced bases. PW represents the time duration of a base being sequenced[23]. Besides the sequenced nucleotides, base modifications would also influence pulse signals. Based on the differences of pulse signals between modified and unmodified bases, methods using SMRT sequencing to detect 5mC and other base modifications have been developed[24]. However, due to the low signal-to-noise ratio, reliable calling of 5mC using early version SMRT data requires high coverage of reads (up to 250×)[15,16]. Recently, the PacBio circular consensus sequencing (CCS) technique was presented[25], in which subreads generated from a circularized template in a single zero-mode waveguide (ZMW) are used to call a consensus sequence (HiFi read) with high accuracy. Using the new CCS technique, Tse *et al.* developed a convolutional neural network (CNN)-based method, called holistic kinetic model (HK model), for genome-wide 5mCpG detection in humans[26]. For a CCS read, HK model

first calculates the mean IPD and PW values of each base after aligning the subreads of the CCS read to genome reference. Then, for each CpG site in the CCS read, HK model organizes the mean IPD values, mean PW values and the sequence context surrounding the CpG site into a feature matrix. At last, HK model feeds the feature matrix into the CNN-based model to get a methylation probability of the CpG site[26]. Experiments on human CCS data showed that, HK model achieved above 90% sensitivity and specificity on 5mCpG detection, and got high correlations with BS-seq on multiple biological samples. However, HK model requires relatively high CCS subread depth (at least 10×) for accurate 5mCpG detection, which limits the insert size in library preparation, further limits the length of CCS subreads. The ability of 5mCpG detection using long PacBio CCS subreads (>10 kb) hasn't been demonstrated.

In this study, we proposed ccsmeth, a deep learning method to detect DNA 5mCpGs from PacBio CCS subreads. ccsmeth uses a deep-learning framework which is composed of bidirectional Gated Recurrent Unit (GRU) and attention neural networks. Compared to HK model on datasets of amplified and M.SssI-treated DNA, ccsmeth achieved higher accuracies for 5mCpG detection. Furthermore, we tested ccsmeth on the CCS data of the human sample HG002 which were sequenced using 15 kb and 24 kb insert libraries[27], and compared ccsmeth with BS-seq and nanopore sequencing on genome-wide 5mCpG detection. The results demonstrated that even using long (>10 kb) PacBio CCS subreads, ccsmeth still got high performances.

Allele-specific methylation (ASM) is closely related to non-coding genomic

elements and gene expression[28]. ASMs occur in both imprinting and non-imprinting regions, which are associated with complex diseases[29,30] and cancers[31]. Recent studies showed that both PacBio SMRT sequencing and nanopore sequencing can be used for haplotype-aware genome assembly[32] and variant calling[33]. However, only nanopore sequencing is now used for accurate haplotype-aware methylation calling and ASM detection[34,35]. Here, with the improved 5mC detection of methhCCS, we further proposed a pipeline to detect haplotype-aware methylation using PacBio CCS. Results on HG002 showed that our proposed pipeline can accurately detect genome-wide allele-specific methylation.

Results

The ccsmeth algorithm for 5mCpG detection

We proposed ccsmeth, a deep learning-based method, to detect CpG methylation from PacBio CCS data. ccsmeth takes raw PacBio CCS subreads and a genome reference as inputs. To infer methylation states of CpGs, the subreads needed to be aligned to the genome reference first. Then, for each CpG, we extracted IPD and PW values of each in a 21-mer sequence context which included the CpG itself in the center from all the aligned subreads. The mean and standard deviation (SD) of IPD and PW values of each base in the 21-mer together with sequence contexts formed a 5×21 feature matrix. Since CpG methylation are mostly symmetric in human[36], as HK model did[26], we constructed two feature matrixes from the subreads in forward and reverse complement strand, respectively, to infer methylation of one CpG (Fig. 1). Different

from HK model, we used two more features: IPD SD and PW SD values.

Recurrent neural network (RNN) is a sequencing process network which has been widely used in base modification detection from long reads[19,20]. In sequencing modeling tasks, attention network can make the model focus on most important "words" through the attention weights. Attention mechanisms have been used to enhance the performance of RNNs in natural language processing[37,38]. Here, in ccsmeth, we designed a deep learning framework, which was constructed by bidirectional GRU[39], Bahdanau attention[40] and fully connected neural networks, to process the feature matrixes of each CpG (Fig. 1b). After processing the feature matrixs, ccsmeth outputted two probability values (P_m and P_{um}) of which the sum was 1 through a Softmax activation function, representing the probabilities of the CpG as methylated and unmethylated, respectively. Finally, the methylation statuses of each CpG at CCS read level and genome level were inferred by ccsmeth (Methods).

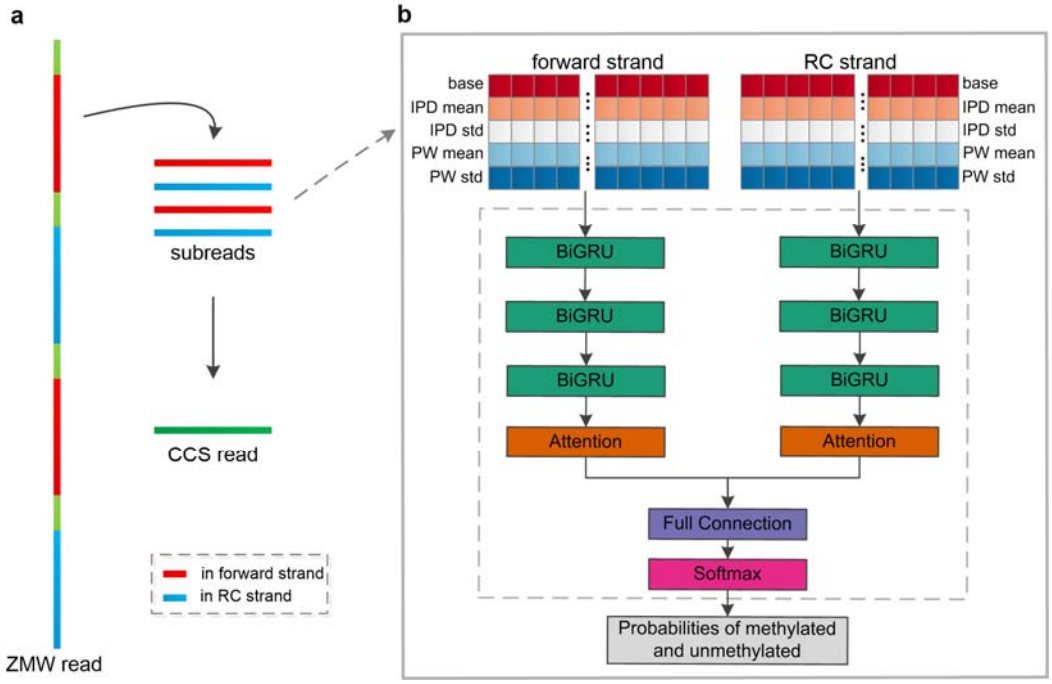

Fig. 1 ccsmeth for 5mCpG detection using PacBio CCS data. **a** PacBio CCS. **b** Model framework of ccsmeth. RC: reverse complement; BiGRU: Bidirectional Gated Recurrent Unit layer; Full Connection: fully connected layer; Softmax: Softmax activation function.

Evalution of ccsmeth using amplified and M.SssI-treated DNA

To evaluate ccsmeth, we first used three groups of CCS datasets sequenced using amplified and M.SssI-treated human DNA from Tse *et al.*[26] (Table 1). In M.SssI-treated DNA, the CpG methyltransferase M.SssI methyalted all CpGs, while amplified DNA which was prepared via whole genome amplification (WGA) contained nearly no methylated bases[26]. As shown in Table 1, M01-03 were M.SssI-treated DNA samples, and W01-03 were amplified DNA samples. The three groups of datasets were sequenced using different version of PacBio sequencers and sequencing kits: PacBio Sequel I sequencer with the Sequel I sequencing kit 3.0 for M01 and W01, PacBio Sequel II sequencer with Sequel II sequencing kit 1.0 for M02 and W02, PacBio Sequel II sequencer with Sequel II sequencing kit 2.0 for M03 and W03. For each group of datasets, we randomly selected 50% methylated reads and unmethylated reads for model training. The remaining 50% methylated and unmethylated reads were used for testing. To demonstrate the ability of ccsmeth on 5mCpG detection, we used the same reads to train and test HK model[26], and then compared the performance of the two methods.

Table 1 Summary of PacBio CCS datasets of amplified and M.SssI-treated DNA used in this study.

sample	sequencing kit	No. of ZMW reads	No. of subreads	mean length of subreads	mean depth of subreads
M01	Sequel I kit 3.0	740,880	44,799,736	336.96	60.47
W01	Sequel I kit 3.0	739,292	43,580,452	343.84	58.95
M02	Sequel II kit 1.0	2,075,537	35,813,589	4,161.45	17.26
W02	Sequel II kit 1.0	2,978,115	42,022,103	4,854.53	14.11
M03	Sequel II kit 2.0	256,865	29,347,179	435.09	114.25
W03	Sequel II kit 2.0	262,608	28,363,026	775.4	108.01

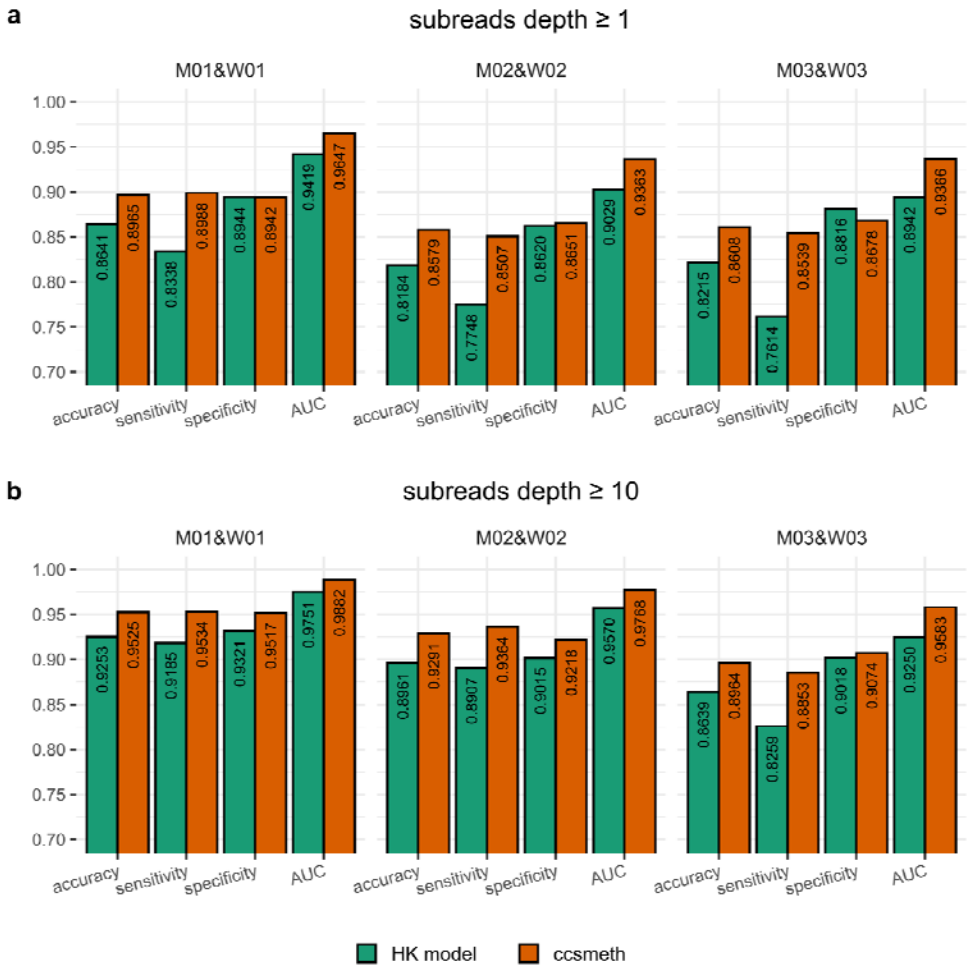

Fig. 2 Comparison of ccsmeth and HK model on 5mCpG detection using datasets of amplified and M.SssI-treated DNA. **a** Evaluating CpG sites with subreads depth ≥1. **b** Evaluating CpG sites with subreads depth ≥10.

As shown in Fig. 2a, when evaluating CpGs with at least 1 subreads depth, while the specificity of HK model was higher than ccsmeth on the Sequel II kit 2.0 dataset (M03 and W03, 0.8816 vs. 0.8678), ccsmeth got higher accuracies, sensitivities and AUCs than HK model on all three datasets. Especially, the accuracies of ccsmeth were above 0.85 and were 3% higher than HK model on all datasets. When further evaluating CpGs with at least 10 subreads depth on all three datasets (Fig. 2b), ccsmeth also got higher performances (higher accuracies, sensitivities, specificities and AUCs) than HK model. On the Sequel II kit 2.0 dataset which had the lowest performance, the accuracy of ccsmeth still got ~0.90 accuracy and ~0.96 AUC, while HK model got 0.8639 accuracy and 0.9250 AUC. The results showed that ccsmeth significantly improved the performance of 5mC detection using PacBio CCS data.

Evalution of ccsmeth against BS-seq and nanopore sequencing

To further test the ability of ccsmeth on 5mCpG detection, we then used three CCS datasets (BC01, NT01, and HepG2) from Tse *et al.*[26] (Table 2). Each of the three datasets has corresponding BS-seq dataset sequenced using the same sample. All three datasets were sequenced using Sequel II kit 1.0. Therefore, we used the models of ccsmeth and HK model trained using datasets M02 and W02 to detect 5mCpG on the three datasets. As Tse *et al.*[26], since the genome-wide coverage of the three datasets was relatively low (Table S1), we only calculated the methylation frequencies of CpGs covered by at least 20 reads using ccsmeth and HK model, and then compared the results with BS-seq. Results showed that ccsmeth got higher Pearson and

Spearman correlations with BS-seq than HK model (Table S1).

Table 2 Summary of PacBio CCS datasets of multiple samples used in this study.

sample	sequencing kit	No. ZMW reads	No. subreads	mean length of subreads	mean depth of subreads
BC01	Sequel II kit 1.0	7,258,611	33,462,507	5,925.68	4.61
NT01	Sequel II kit 1.0	6,000,227	69,707,261	4,940.29	11.62
HepG2	Sequel II kit 1.0	6,731,204	45,596,758	7,376.23	6.77
HG002-15kb	Sequel II kit 2.0	12,355,008	86,390,109	14,584.51	6.99
HG002-24kb	Sequel II kit 2.0	11,307,763	51,094,330	21,854.12	4.52

The read length of all CCS datasets from Tse *et al.*[26] was less than 10kb (Table 2). Therefore, the ability for 5mCpG detection using long PacBio CCS reads was not assessed. From Baid *et al.*[27], we further got two PacBio CCS datasets of the human HG002 sample: dataset HG002-15kb sequenced using 15 kb insert library, and dataset HG002-24kb sequenced using 24 kb insert library (Table 2). Besides the two CCS datasets of HG002, we also got BS-seq and nanopore sequencing datasets of HG002 from ONT Open Datasets (Methods). For model training and evaluation, we used the result of BS-seq which was processed by Bismark[41] as gold standard (Methods). One million ZMW reads were randomly selected from HG002-15kb for training models of ccsmeth and HK model, while the remaining reads of HG002-15kb and all reads of HG002-24kb were used for testing.

We first accessed the performance of ccsmeth for 5mCpG detection on the two HG002 datasets at read level. As shown in Fig. 3a, ccsmeth achieved higher performances than HK model. On dataset HG002-15kb, ccsmeth got 0.8953 accuracy and 0.9631 AUC, which were ~3% and 2% higher than HK model, respectively. On dataset HG002-24kb, ccsmeth also got higher accuracy and AUC, which were 0.8773 and 0.9517, respectively. ccsmeth outputs two probability values P_m and P_{um} for a

targeted CpG in a read, representing the probability of the CpG being methylated and unmethylated, respectively (Methods). By requiring the value of $|P_m\text{-}P_{um}|$ to reach a threshold Δ_p to make a call for a CpG, the performance of ccsmeth can be further improved at the cost of making fewer calls (Fig. S1a and b). To balancing the performance of ccsmeth and the number of made calls, we set Δ_p as 0.5. With Δ_p=0.5, the accuracy and AUC of ccsmeth on dataset HG002-15kb can be improved to ~0.945 and ~0.98 (Fig. 3a), respectively, with 21.3% ambiguous calls being discarded. On dataset HG002-24kb, the accuracy and AUC were improved to ~0.935 and >0.97 (Fig. 3a), respectively, while 24.4% ambiguous calls were discarded.

Using the HG002 datasets, we then compared per-site methylation frequencies of ccsmeth with BS-seq and nanopore sequencing at genome level. We used the methylation frequencies of CpGs which were covered by at least 5 reads for comparison. As shown in Fig. 3b, using dataset HG002-15kb, ccsmeth got 0.8887 and 0.8873 Pearson correlations with BS-seq and nanopore sequencing, respectively. With dataset HG002-24kb, ccsmeth also got >0.86 Pearson correlations. The 5mCpG detection of ccsmeth on datasets HG002-15kb and HG002-24kb also showed great consistent with a Pearson correlation value of >0.95. Compared to HK model, the Pearson correlations and Spearman correlations of ccsmeth with BS-seq and nanopore sequencing were higher: >3% higher for dataset HG002-15kb and >2% higher for dataset HG002-24kb (Table S2). ccsmeth also got lower root mean square errors (RMSEs) than HK model, except when comparing with nanopore sequencing on dataset HG002-24kb (0.1847 vs. 0.1827).

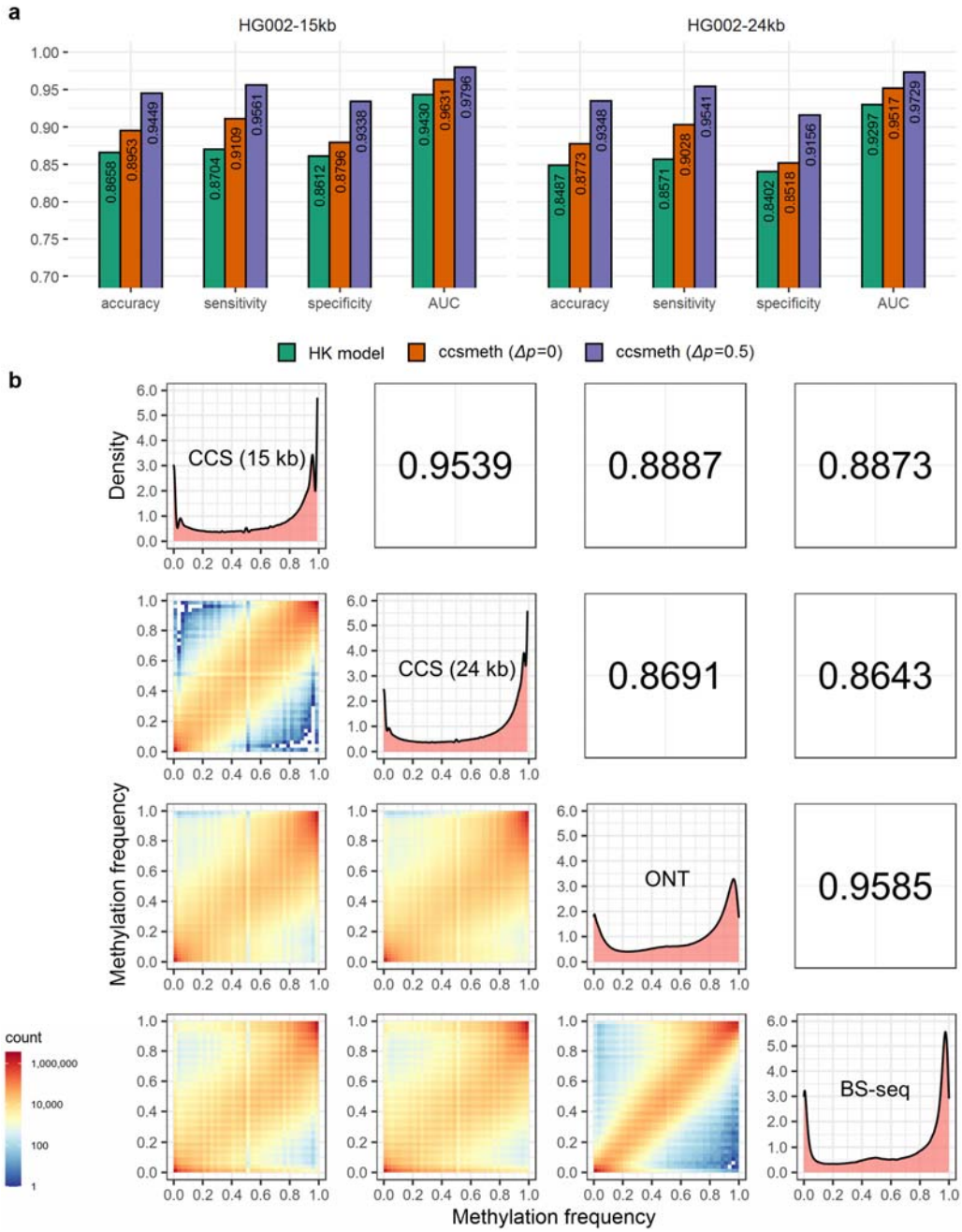

Fig. 3 Evaluation of ccsmeth on HG002 datasets. **a** Comparison ccsmeth and HK model at read level. Δ_p: probability threshold. **b** Comparing ccsmeth with BS-seq and nanopore sequening at genome level. Values in upper triangle are Pearson correlations. CCS: PacBio CCS sequencing; ONT: nanopore sequencing; BS-seq: bisulfite sequencing.

Haplotype-aware methylation calling using PacBio CCS data

With the improvements in read quality, yield, and base modification, long read sequencing has the potential to allow genome assembly, detection of structural variants (SVs), single nucleotide variants (SNVs) and methylation using a single sequencing run. Using nanopore sequencing, Akbari *et al.* proposed a tool called NanoMethPhase for haplotype-aware methylation calling and ASM detection[34]. Similar to the framework of NanoMethPhase, we proposed a pipeline to call haplotype-aware methylation and detect ASMs using PacBio CCS data (Fig. 4a, Methods). In this pipeline, ccsmeth was used to call 5mCpG methylation in each read and Clair3[42] was used to call SNVs. The called SNVs was used to phase reads of two haplotypes by WhatsHap[43]. Finally, as in NanoMethPhase, the R package DSS[44] was used to call differentially methylated regions (DMRs) from the reads of the two haplotypes.

We used two approaches to access this pipeline using the HG002 PacBio CCS data. First, since DNA methylation of CpGs in imprinting control regions (ICRs) mediates parental allele-specific expression of imprinted genes[45], we investigated haplotype-aware methylation status in ICRs of HG002 as Akbari *et al.*[34] did. We got 43 well-known ICRs and 48 novel ICRs. The 48 novel ICRs were collected from Court *et al.*[46] and Joshi *et al.*[47]. After phasing reads to two haplotypes using WhatsHap, we calculated the methylation levels of each ICR at the two haplotypes using ccsmeth (Methods). Then we compared the differences of methylation levels at the two haplotypes for each ICR (Fig. 4b). Among the tested ICRs, >90% known

ICRs and ~50% novel ICRs got >0.2 methylation level difference. We further tested these ICRs using nanopore data and a similar pipeline for comparison (Note S1). As shown in Figure S2, the methylation level differences of the ICRs got from PacbBio CCS data were quite consistent with those from nanopore data.

We got 9994 DMRs in HG002 using the proposed pipeline and the PacBio CCS data. Previous studies have shown that most imprinted genes are clustered at regions which can span up to approximately 4 Mb with their ICRs[34,48]. Thus, we then assessed the called DMRs by investigating their distance with imprinting genes. We got 107 imprinted, 103 predicted, 14 conflicting, and 6 provisional genes from Akbari *et al.*[34]. ~45.6% of the 9994 DMRs were located at upstream and downstream <4 Mb regions of the 230 genes. Moreover, 227 of the 230 genes have at least one DMR in its upstream and downstream <1 Mb regions (Fig. 4c and d).

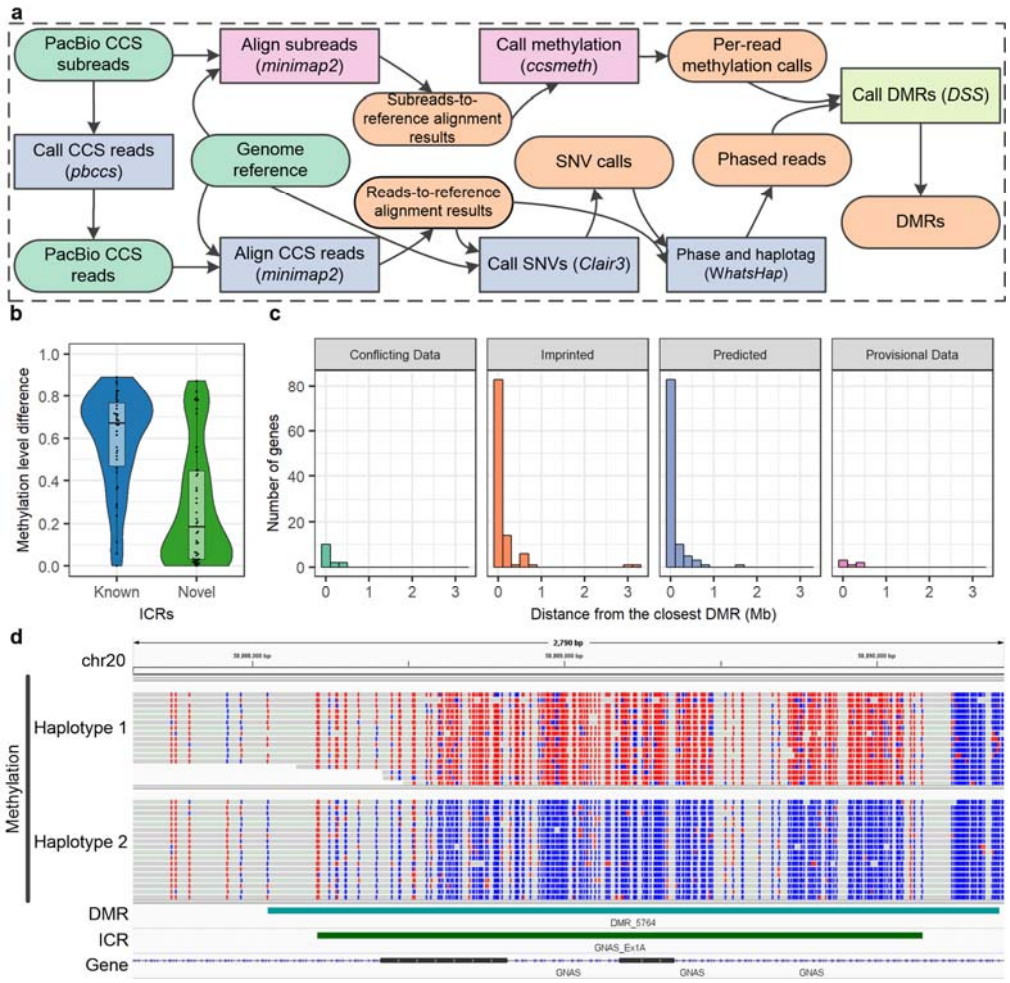

Fig. 4 Haplotype-aware methylation calling using CCS data. **a** Pipeline of haplotype-aware methylation calling with ccsmeth and other tools. **b** Distribution of methylation level differences of ICRs between two haplotypes of HG002. Boxes inside the violinplots are boxplots, dots are values of the methylation level differences. **c** Distribution of the number of known, predicted, conflicting, and provisional imprinted genes in terms of distance from the closest DMR of each gene. **d** Screenshot of Integrative Genomics Viewer on a DMR of HG002 near the imprinted gene *GNAS* (chr20:58,887,618-58,890,406).

Discussion

Due to the high accuracy and long reads, SMRT sequencing has become more and more widely used in genome related research. However, compared to Nanopore sequencing, the application of SMRT sequencing on DNA 5mC detection was not fully studied before. Recently, Tse *et al.* developed HK model, a convolutional neural network (CNN)-based method, to detect 5mCpG in humans[26], which demonstrated that SMRT sequencing can detect 5mC accurately under acceptable coverage of reads.

In this study, we proposed ccsmeth to detect 5mCpGs from PacBio SMRT CCS reads. Using BiGRU and attention mechanism, ccsmeth significantly improved the performance of SMRT sequencing for 5mCpG detection. We also proposed a pipeline to call haplotype-aware methylation using CCS data. It has been demonstrated that PacBio SMRT sequencing can be used to detect haplotype-resolved SNVs and structural variations (SVs)[35]. Our results of the proposed pipeline on HG002 further indicated that haplotype-aware SNVs, SVs, and methylation may be accurately co-analyzed using only PacBio CCS data.

We only tested ccsmeth on detection of 5mCpGs in this study. With more datasets generated, detection of 5mC in non-CpGs, and other types of DNA methylation (such as 5hmC) can be studied in future research. In ccsmeth, considering that 5mC in CpGs are usually symmetric, we combined the features of sequences on both DNA strands for each targeted CpG for more accurate prediction. However, this cannot be used for 5mC detection in non-CpGs or detection of other types of methylation. Hence, methods may also need to be re-designed to detect more types of methylation.

Methods

PacBio CCS data of M.SssI-treated and amplified DNA. We downloaded three PacBio CCS datasets of human from Tse *et al.*[26]. Each of the three datasets contains two groups of reads: the methylated reads sequenced using M.SssI-treated DNA and the unmethylated reads sequenced using amplified DNA. As shown in Table 1, reads of sample M01 (M.SssI-treated DNA) and sample W01 (amplified DNA) were sequenced on the PacBio Sequel I sequencer with the Sequel I sequencing kit 3.0. Reads of sample M02 and W02 were sequenced on the PacBio Sequel II sequencer with Sequel II sequencing kit 1.0. Reads of sample M03 and W03 were sequenced on the PacBio Sequel II sequencer with Sequel II sequencing kit 2.0.

Datasets of multiple human biological samples. From Tse *et al.*[26], We downloaded PacBio CCS data of three human biological samples: Buffy coat (BC01), adjacent normal tissue (NT01), and HCC cell line (HepG2) (Table 3). All the three datasets were sequenced using PacBio Sequel II sequencer with Sequel II sequencing kit 1.0. We also got BS-seq data of the three samples from Tse *et al.* for evaluation.

We downloaded two PacBio CCS datasets of human HG002 sample from Baid *et al.*[27]. One of the two datasets was sequenced using library with insert size of 15 kb (HG002-15kb), another was sequenced using library with insert size of 24 kb (HG002-24kb) (Table 3). Both the two datasets were sequenced on a Sequel II sequencer with sequencing kit 2.0. There were >1.2 million ZMW reads of HG002-15kb, and >1.1 million ZMW reads of HG002-24kb. Using the ZMW reads and the PacBio official tool pbccs (v6.2.0, https://github.com/PacificBiosciences/ccs),

we generated ~26× mean genome coverage of CCS/HiFi reads for HG002-15kb and ~28× coverage of CCS reads for HG002-24kb. From the datasets, we randomly selected 1 million ZMW reads of HG002-15kb for model training. The remaining reads of HG002-15kb and all the reads of HG002-24kb were used for evaluation. To train and evaluate ccsmeth, we got BS-seq and nanopore sequencing data of HG002 as standard from ONT Open Datasets (https://labs.epi2me.io/dataindex/). There were ~100× coverage of BS-seq reads. We basecalled raw nanopore reads using Guppy (version 4.2.2+effbaf8) and got ~60× coverage of nanopore reads for downstream analysis.

Framework of ccsmeth. ccsmeth contains the following three steps for methylation detection (Fig.1):

(1) Alignment. Raw CCS subreads should be aligned to genome reference first. In ccsmeth, we provided two popular tools, BWA-MEM[49] or minimap2[50], for alignment. For the CCS data from Tse *et al.*[26], we used BEA-MEM for alignment. For the CCS data of HG002 which have long reads, we used minimap2 for alignment. We used hg38/GRCh38 as the human genome reference.

(2) Feature extraction. After alignment, we first normalized the IPD and PW values of each subread using Z-score normalization. Then we extracted the normalized IPD and PW values of each base from the corresponding subreads. For each targeted CpG, we extracted 21-mer surrounding the base. Then we calculated the mean and std of IPD and PW values of each base in the 21-mer. Therefore, with sequence contexts, a 5×21 feature matrix was formed for each CpG. Because of the symmetric methylation

feature of CpG sites, we further combined two feature matrixes which were from both DNA strands for prediction (Fig. 1).

(3) Methylation calling. Each of the two feature matrixes of a CpG in a read was fed into a deep neural network of ccsmeth, which contained three bidirectional Gated Recurrent Unit (GRU) layers[39] and one Bahdanau attention layer[40] (Fig.1, Note S2). The outputs from the two attention layers were processed by a full connection layer and then by a Softmax layer. The Softmax layer outputted two probabilities P_m and P_{um}. P_m represented the probability of the CpG being methylated, while P_{um} represented the probability of the CpG being unmethylated. The sum of P_m and P_m is 1. Using P_m and P_{um}, the methylation status of the CpG at read level can be predicted: If $P_m > P_{um}$, the CpG is predicted as methylated, otherwise is predicted as unmethylated. Furthermore, using the difference of P_m and P_{um} ($|P_m - P_{um}|$), the ambiguous calls of ccsmeth can be discarded to improve the performance for 5mCpG prediction (Fig. S1). At last, by counting the number of reads where the CpG was called methylated and the total number of reads mapped to the CpG, the methylation frequency of the CpG at genome level can be calculated.

Model training of ccsmeth. For the datasets of M.SssI-treated and amplified DNA, we extracted positive (methylated) and negative (methylated) samples from reads of M.SssI-treated and amplilfied DNA, respectively. For the HG002 datasets, we first selected cytosines with one and zero methylation frequency according to the results of BS-seq. Then we extracted the sites from the corresponding positions in the reads which were mapped to the cytosines with one and zero methylation frequency as

positive and negative samples, respectively. After collecting samples for training, we split the total samples at a ratio of 99:1 as training dataset and validation dataset. The model parameters were learned on the training dataset by minimizing the loss calculated by cross-entropy with a batch size of 512 and an initial learning rate of 0.001. The learning rate was adopted by Adam optimizer[51] and decayed by a factor of 0.1 after every epoch. The parameter betas in Adam optimizer were set to (0.9, 0.999). We used two strategies to prevent overfitting. First, we added dropout layer in GRU layers and the fully connected layer. We set dropout probability to 0.5 at each dropout layer. Second, we used early stopping[52] during training. We set at least 5 epochs and at most 10 epochs for each training. The model parameters with the current best performance on validation dataset are saved in every epoch. During epoch 6 to 10, if the best performance of the current epoch decreases, we stop the training process.

Evaluation of ccsmeth. We first evaluated ccsmeth at read level using the three datasets of M.SssI-treated and amplified human DNA. For each dataset, we randomly selected 50% methylated reads and 50% unmethylated reads for model training, while the remaining 50% methylated and unmethylated reads were used for evaluation. We calculated accuracy, sensitivity, specificity, and area under the curve (AUC) based on the prediction of randomly selected 100,000 positive samples and 100,000 negative samples from the reads used for evaluation. We repeated the subsampling evaluation 5 times.

We then used the datasets of multiple biological samples to evaluate ccsmeth. For comparison, we used Bismark[41] (v0.22.3) to process the BS-seq data of the biological

samples (BC01, NT01, HepG2, and HG002). We used deepsignal2, an improved version of deepsignal[20], to process the nanopore data of HG002. Based on the results of BS-seq data of HG002, we selected cytosines with one and zero methylation frequency as methylated (positive) and unmethylated (negative) sites. Then we used the sites in the reads which were mapped to the methylated and unmethylated sites to evaluate ccsmeth at read level.

To evaluate ccsmeth at genome level, we compared per-site methylation frequencies predicted by ccsmeth with the results of BS-seq and nanopore sequencing on all biological samples. Using the per-site methylation frequencies, Pearson correlation (r), the coefficient of determination (r^2), Spearman correlation (ρ), and root mean square error (RMSE) were calculated.

Pipeline for haplotype-aware methylation calling using PacBio CCS data. In the proposed pipeline to call haplotype-aware methylation, we first aligned the PacBio CCS subreads to genome reference using minimap2[50] (version 2.22-r1101) and then used ccsmeth for methylation calling. For variant calling, the CCS reads also needed to be aligned to genome reference by minimap2 first. Then, we used Clair3[42] (v0.1) with "*hifi*" model to call variants. The called "PASS" SNVs were then used by WhatsHap[43] (version 1.2) to phase and tag the reads to two haplotypes. At last, the reads of the two haplotypes together with the methylation calling results were used to call DMRs by DSS[44] (version 2.43.2).

To measure the methylation level of ICRs in haplotypes, we counted the total number of CpGs and the number of CpGs which were predicted as methylated by

ccsmeth in the phased reads mapped to each ICR. Then the methylation level of an ICR in one haplotype was calculated as the number of methylated CpGs divided by the total number of CpGs.

Data availability

The CCS datasets of M.SssI-treated and amplified DNA (M01-03, W01-03), the CCS and BS-seq datasets of three biological samples (BC01, NT01, and HepG2) are available from Tse *et al.*[26]. The CCS reads of HG002 are available from Google Cloud[27] [https://console.cloud.google.com/storage/browser/brain-genomics-public/research/deepconsensus/publication/sequencing]. Raw nanopore reads of HG002 are available from flowcell PAG07165 of ONT Open Datasets [https://labs.epi2me.io/gm24385_2020.11/]. The BS-seq reads of HG002 are also available from ONT Open Datasets [https://labs.epi2me.io/gm24385-5mc/].

Code availability

ccsmeth is publicly available at GitHub [https://github.com/PengNi/ccsmeth].

Acknowledgments

This work makes use of program (for HK model) and data generated by The Chinese University of Hong Kong (CUHK) Department of Chemical Pathology, as reported by Tse *et al.*, of whom we are very thankful, in Proc Natl Acad Sci USA 2021; 118(5): e2019768118 (https://doi.org/10.1073/pnas.2019768118). We thank ONT for making

their Nanopore and BS-seq data of HG002 publicly available. We also thank Baid *et al.* for sharing their CCS raw reads of HG002 publicly. This work was carried out in part using computing resources at the High Performance Computing Center of Central South University. This work was supported in part by the National Natural Science Foundation of China under Grants (Nos. U1909208, 61732009 and 61832019), 111 Project (No. B18059), Hunan Provincial Science and Technology Program (No. 2018wk4001).

References

1 Breiling, A. & Lyko, F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics & chromatin* **8**, 1-9 (2015).

2 Robertson, K. D. DNA methylation and human disease. *Nature Reviews Genetics* **6**, 597-610 (2005).

3 Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature Reviews Genetics* **14**, 204-220 (2013).

4 Gonzalo, S. Epigenetic alterations in aging. *Journal of applied physiology* **109**, 586-597 (2010).

5 Miura, F., Enomoto, Y., Dairiki, R. & Ito, T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic acids research* **40**, e136-e136 (2012).

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

7 Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences* **89**, 1827-1831 (1992).

8 Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13**, 36-46 (2012).

9 Liu, Y. *et al.* Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nature biotechnology* **37**, 424-429 (2019).

10 Vaisvila, R. *et al.* Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Research* (2021).

11 Yang, Y. & Scott, S. A. in *Functional Genomics* 125-134 (Springer, 2017).

12 Liu, Y. *et al.* Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. *Genome biology* **21**, 1-9 (2020).

13 Sun, Z. *et al.* Nondestructive enzymatic deamination enables single-molecule

long-read amplicon sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Genome research* **31**, 291-300 (2021).

14 Sakamoto, Y. *et al.* Long-read whole-genome methylation patterning using enzymatic base conversion and nanopore sequencing. *Nucleic Acids Research* (2021).

15 Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome biology* **21**, 1-16 (2020).

16 Gouil, Q. & Keniry, A. Latest techniques to study DNA methylation. *Essays in biochemistry* **63**, 639-648 (2019).

17 Stoiber, M. *et al.* De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv*, 094672, doi:10.1101/094672 (2017).

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

19 ONT. *Megalodon*, <https://github.com/nanoporetech/megalodon> (2022)

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

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

22 Liu, Y. *et al.* DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biology* **22**, 295, doi:10.1186/s13059-021-02510-z (2021).

23 Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods* **7**, 461-465 (2010).

24 Feng, Z. *et al.* Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. *PLoS computational biology* **9**, e1002935 (2013).

25 Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* **37**, 1155-1162, doi:10.1038/s41587-019-0217-9 (2019).

26 Tse, O. O. *et al.* Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proceedings of the National Academy of Sciences* **118** (2021).

27 Baid, G. *et al.* DeepConsensus: Gap-Aware Sequence Transformers for Sequence Correction. *bioRxiv* (2021).

28 Chess, A. Monoallelic expression of protocadherin genes. *Nature genetics* **37**, 120-121 (2005).

29 Benton, M. C. *et al.* Genome-wide allele-specific methylation is enriched at gene regulatory regions in a multi-generation pedigree from the Norfolk Island isolate. *Epigenetics & Chromatin* **12**, 60, doi:10.1186/s13072-019-0304-7 (2019).

30 Plongthongkum, N., Diep, D. H. & Zhang, K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nature Reviews Genetics* **15**, 647-661 (2014).

31 Jelinic, P. & Shaw, P. Loss of imprinting and cancer. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* **211**, 261-268 (2007).

32 Luo, X., Kang, X. & Schönhuth, A. phasebook: haplotype-aware de novo assembly of diploid genomes from long reads. *Genome Biology* **22**, 299, doi:10.1186/s13059-021-02512-x (2021).

33 Shafin, K. *et al.* Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods* **18**, 1322-1332, doi:10.1038/s41592-021-01299-w (2021).

34 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).

35 Mahmoud, M., Doddapaneni, H., Timp, W. & Sedlazeck, F. J. PRINCESS: comprehensive detection of haplotype resolved SNVs, SVs, and methylation. *Genome Biology* **22**, 268, doi:10.1186/s13059-021-02486-w (2021).

36 Arand, J. *et al.* In Vivo Control of CpG and Non-CpG DNA Methylation by DNA Methyltransferases. *PLOS Genetics* **8**, e1002750, doi:10.1371/journal.pgen.1002750 (2012).

37 Yang, Z. *et al.* in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies.* 1480-1489.

38 Zhou, P. *et al.* in *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers).* 207-212.

39 Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

40 Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

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

42 Zheng, Z. *et al.* Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *bioRxiv*, 2021.2012.2029.474431, doi:10.1101/2021.12.29.474431 (2021).

43 Patterson, M. *et al.* WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology* **22**, 498-509 (2015).

44 Park, Y. & Wu, H. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics* **32**, 1446-1453 (2016).

45 Maupetit-Méhouas, S. *et al.* Imprinting control regions (ICRs) are marked by mono-allelic bivalent chromatin when transcriptionally inactive. *Nucleic acids research* **44**, 621-635, doi:10.1093/nar/gkv960 (2016).

46 Court, F. *et al.* Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome research* **24**, 554-569 (2014).

47 Joshi, R. S. *et al.* DNA methylation profiling of uniparental disomy subjects provides a map of parental epigenetic bias in the human genome. *The American Journal of Human Genetics* **99**, 555-566 (2016).

48 da Rocha, S. T. & Gendrel, A.-V. The influence of DNA methylation on monoallelic expression. *Essays in biochemistry* **63**, 663-676 (2019).

49 Li, H. Aligning sequence reads, clone sequences and assembly contigs with

BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).

50 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).

51 Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

52 Prechelt, L. in *Neural Networks: Tricks of the trade* 55-69 (Springer, 1998).