*Bioinformatics*, 35(22), 2019, 4586–4595
doi: 10.1093/bioinformatics/btz276
Advance Access Publication Date: 17 April 2019
Original Paper





Genome analysis

# DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning

**Peng Ni[1,†], Neng Huang[1,†], Zhi Zhang[1], De-Peng Wang[2], Fan Liang[2], Yu Miao[2], Chuan-Le Xiao[3], Feng Luo[4,*] and Jianxin Wang [1,*]**

[1]School of Information Science and Engineering, Central South University, Changsha 410083, China, [2]GrandOmics Biosciences, Beijing 102206, China, [3]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China and [4]School of Computing, Clemson University, Clemson, SC 29634, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Arne Elofsson

Received on January 1, 2019; revised on April 6, 2019; editorial decision on April 8, 2019; accepted on April 11, 2019

## Abstract

**Motivation:** The Oxford Nanopore sequencing enables to directly detect methylation states of bases in DNA from reads without extra laboratory techniques. Novel computational methods are required to improve the accuracy and robustness of DNA methylation state prediction using Nanopore reads.

**Results:** In this study, we develop DeepSignal, a deep learning method to detect DNA methylation states from Nanopore sequencing reads. Testing on Nanopore reads of *Homo sapiens* (*H. sapiens*), *Escherichia coli* (*E. coli*) and pUC19 shows that DeepSignal can achieve higher performance at both read level and genome level on detecting 6 mA and 5mC methylation states comparing to previous hidden Markov model (HMM) based methods. DeepSignal achieves similar performance cross different DNA methylation bases, different DNA methylation motifs and both singleton and mixed DNA CpG. Moreover, DeepSignal requires much lower coverage than those required by HMM and statistics based methods. DeepSignal can achieve 90% above accuracy for detecting 5mC and 6 mA using only 2× coverage of reads. Furthermore, for DNA CpG methylation state prediction, DeepSignal achieves 90% correlation with bisulfite sequencing using just 20× coverage of reads, which is much better than HMM based methods. Especially, DeepSignal can predict methylation states of 5% more DNA CpGs that previously cannot be predicted by bisulfite sequencing. DeepSignal can be a robust and accurate method for detecting methylation states of DNA bases.

**Availability and implementation:** DeepSignal is publicly available at https://github.com/bioinfomaticsCSU/deepsignal.

**Contact:** luofeng@clemson.edu or jxwang@mail.csu.edu.cn

**Supplementary information:** Supplementary data are available at *bioinformatics* online.

## 1 Introduction

DNA methylation, as a crucial form of epigenetic marks, plays important roles in a number of key biological processes (Bergman and Cedar, 2013; Schübeler, 2015). N6-methyladenine (6 mA) and 5-methylcytosine (5mC) are the two most prevalent and well-studied DNA base methylations. 5mC usually plays a role in

**Table 1.** Methods for DNA methylation detection from Nanopore sequencing reads

| Type | Method | Technique | Reference |
|---|---|---|---|
| Model based | Nanopolish | HMM | Simpson *et al.*, 2017 |
| | signalAlign | HMM+HDP | Rand *et al.*, 2017 |
| | mCaller | Neural network random forest naive Bayes logistic regression | McIntyre *et al.*, 2017 |
| Statistics based | nanoraw | Mann-Whitney U test | Stoiber *et al.*, 2016 |
| | NanoMod | Kolmogorov-Smirnov test | Liu *et al.*, 2019 |

embryonic development (Smith and Meissner, 2013), atherosclerosis (Lund *et al.*, 2004), aging and diseases (Gonzalo, 2010); and 6 mA is important in transcriptional regulation (Yue *et al.*, 2015), cancer development (Xiao *et al.*, 2018) and neuro development (Yao *et al.*, 2018).

Recently, single molecule sequencing technologies, such as PacBio single molecule real time (SMRT) sequencing and Nanopore sequencing, are demonstrated to be able to detect the DNA methylation marks directly. Both technologies distinguish modified bases from standard nucleotide bases based on their distinctive signals. For PacBio SMRT sequencing, base modification would affect DNA polymerase kinetics, and then can be detected through different inter-pulse duration (Flusberg *et al.*, 2010). However, the accuracy of SMRT sequencing for detecting DNA methylation is heavily affected by the sequence coverage (Davis *et al.*, 2013; Zhu *et al.*, 2018). Meanwhile, it also has been found that electrical signals in Nanopore sequencing are sensitive to epigenetic changes in the nucleotides (Laszlo *et al.*, 2013; Schatz, 2017; Schreiber *et al.*, 2013). Several studies have demonstrated that Nanopore sequencing can be used to detect DNA methylation.

Both model based and statistics based methods have been developed to identify DNA methylation from Nanopore sequencing reads (Table 1). For a targeted DNA base, model based methods first predict the methylation state of targeted base in each mapped read (Supplementary Fig. S1A). Then, the predictions of the targeted base in all mapped reads are summarized to get the methylation state of targeted base at genome level. The prediction at genome level can be either a binary methylation call or a methylation frequency (i.e. number of mapped reads predicted as methylated/number of total mapped reads). Simpson *et al.* (2017) proposed a hidden Markov model (HMM) based approach, which is included in their nanopolish tool, to detect DNA 5mC in CpG from events of Nanopore reads. The HMM model trained from *Escherichia coli* data can detect DNA CpG methylation in *Homo sapiens* Nanopore reads with 87% accuracy at read level. Because of limited training data, nanopolish was neither able to identify non-CpG methylation nor distinguish k-mers which contain both methylated and unmethylated CpGs. Rand *et al.* (2017) proposed a tool called signalAlign, which used HMM with hierarchical Dirichlet process (HDP) to classify 5mC at the inner cytosine of CCWGG motifs (W represents A or T) and 6 mA at GATC motifs from events of Nanopore reads. singalAlign achieved 86–95% accuracies at genome level for the Nanopore R9 data of pUC19 and *E. coli*. McIntyre *et al.* (2017) used four kinds of classifiers (neural network, random forest, naive Bayes and logistic regression) to detect DNA 6 mA in mouse, *E. coli* and Lambda phage data. They achieved 84% accuracy at read level and 94% accuracy at genome level under 15× or higher coverage.

To identify modified DNA bases, statistics based methods require two groups of reads: one group of reads from a native DNA and another group of reads from a matched amplified DNA. By testing the significance of difference between the two groups of reads, the methylation state of a targeted DNA base at genome level is predicted (Supplementary Fig. S1B). Stoiber *et al.* (2016) adopted Mann-Whitney U test to detect controlled DNA methylation in *E. coli*. They achieved 0.62–0.97 genome level AUCs for seven different 5mC and 6 mA motifs. Liu *et al.* (2019) used Kolmogorov-Smirnov test in their NanoMod tool to identify modified DNA bases. Testing on *E. coli* methylation data, NanoMod have above 70% precision with 50% recall at genome level. Statistics based methods don't need training datasets to detect base modifications. However, to identify the modified bases in a native DNA, statistics based methods require a matched amplified DNA as a control sample. Furthermore, comparing to model based methods, statistics based methods achieved less accuracy (Liu *et al.*, 2019; Stoiber *et al.*, 2016).

Here, we present a deep learning method, called DeepSignal, to predict the methylation states of DNA bases using Nanopore sequencing read. DeepSignal employs two modules to construct features from raw electrical signals of Nanopore reads (Fig. 1). First, the *signal* feature module in DeepSignal uses convolutional neural network (CNN) to construct features directly from raw electrical signals around methylated base. Second, the *sequence* feature module in DeepSignal uses bidirectional recurrent neural network (BRNN) to construct features from sequences of signal information. Then, features built from two modules are concatenated and fed into a fully connected neural network to predict the methylation states. We first evaluate DeepSignal on controlled DNA 5mC and 6 mA data. We show that DeepSignal achieves higher accuracy at both read level and genome level and requires less coverage of reads than previous HMM based methods do. Then, we compare DeepSignal to bisulfite sequencing based method using data from *H. sapiens* HX1 and NA12878 samples. DeepSignal can achieve above 0.9 Pearson correlation with bisulfite sequencing data of both samples with using only 20× coverage of reads. Furthermore, DeepSignal can achieve higher coverage of DNA CpG sites than bisulfite sequencing does.

## 2 Materials and methods

### 2.1 Data

#### 2.1.1 Nanopore sequencing data

**CpG methylation (5mC) data.** The Nanopore reads of CpG methyltransferase M.SssI-treated (methylated) and PCR-amplified (unmethylated) *E. coli* K12 ER2925 and *H. sapiens* NA12878 DNA previously used by Simpson *et al.* (2017) are downloaded from the European Nucleotide Archive under accession PRJEB13021. The dataset contains reads from multiple sequencing runs using Nanopore R7.3 or R9 flow cells with 2D sequencing technology. During 2D sequencing, double strands of a DNA molecule are linked by a hairpin adaptor, and then sequenced by Nanopore. Thus, 2D technology allows to obtain two reads for one DNA molecule. Since there are no raw signals provided in the R7.3 reads, only the R9 2D reads with raw signals are used in our study.

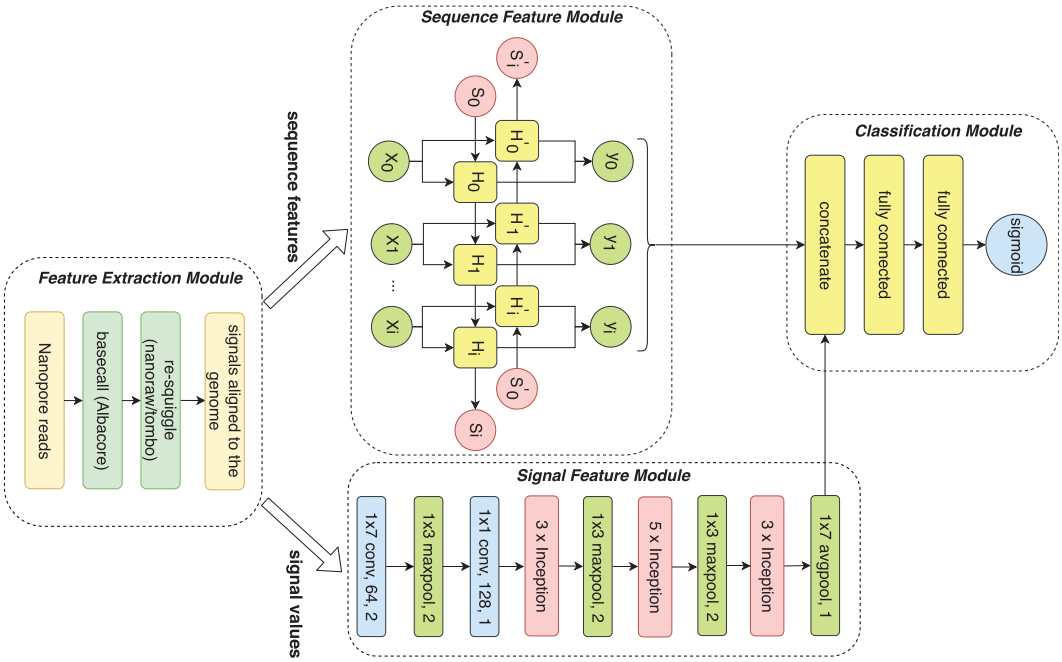


**Fig. 1.** DeepSignal model for detecting DNA methylation states using Nanopore reads [In *Signal Feature Module*, (1 × 7 conv, 64, 2) represents a convolution layer with a (1, 7) kernel size, 64 output channels and a stride of 2. (1 × 3 maxpool, 2) represents a maxpooling layer with a (1, 3) kernel size and a stride of 2. (3× inception) represents three tandem inception modules (Szegedy *et al.*, 2015). (5× inception) represents five tandem inception modules. (1 × 7 avgpool, 1) represents an average pooling layer with a (1, 7) kernel size and a stride of 1. In CNN, a kernel is a convolution matrix. The stride is the step size each time a convolution sliding matrix moves)

The 30× R9.4 1D Nanopore reads of NA12878 DNA are downloaded from the European Nucleotide Archive under accession PRJEB23027 (Jain *et al.*, 2018). About 50× R9.4 1D Nanopore reads of HX1 native DNA are generated in house by GrandOmics. We also generate 5× whole genome amplification (WGA) reads of HX1 as control sample.

**GATC methylation (6 mA) and CCWGG methylation (5mC) Data.** The Nanopore R9 2 D reads of methylated and unmethylated pUC19 plasmid DNA (both 5mC and 6 mA data) previously used by Rand *et al.* (2017) are downloaded from NCBI under accession SRP098631. We select the reads of which the 2D sequences cover more than 2600 bp of the pUC19 DNA reference. After the alignment using BWA-MEM (Li, 2013), we obtain 12 844 reads of methylated genome (gDNA) and 18 892 reads of unmethylated genome (pcrDNA).

### 2.1.2 Bisulfite sequencing data

The analysis results of bisulfite sequencing of *H. sapiens* NA12878 (55×) is downloaded from ENCODE (ENCFF835NTC) (Consortium *et al.*, 2012). Two technical replicates (32× and 34×) of bisulfite sequencing data of HX1 are generated by GrandOmics. We analyze the HX1 bisulfite sequencing using Bismark (Krueger and Andrews, 2011) with the GRCh38.p5 *H. sapiens* genome reference. For each CpG site, Bismark outputs a methylation call for each of its mapped reads, and then calculates the methylation frequency.

To train our model, we select high confidence DNA CpG sites based on bisulfite sequencing data. A CpG is said to be methylated with high confidence if the CpG has 100% methylation frequency and is covered with at least five reads. And a CpG is said to be unmethylated if it has at least five mapped reads and the methylation frequency is zero. We only choose CpGs with consistent methylation states on both the forward and the backward strand as high-confidence sites. For HX1, we take the union of high-confidence sites from two replicates as our final high-confidence set. For NA12878, the selected high-confidence sites of chrY are removed. In total, 12 784 850 methylated CpGs and 976 586 unmethylated CpGs are selected from HX1 bisulfite sequencing data. 5 408 142 methylated CpGs and 4 657 638 unmethylated CpGs are selected from NA12878 bisulfite sequencing data.

## 2.2 DeepSignal model

As shown in Figure 1, DeepSignal is made up of four modules. The first module extracts raw signals from Nanopore reads and prepares the input data for feature construction modules. Two modules are designed to construct different features: the *signal* feature module that constructs features from raw electrical signals around modified base using CNN, and the *sequence* feature module that constructs features from the sequence of signal information around modified base using BRNN. Finally, a classification module of neural network takes features built from two feature modules and predicts DNA methylation states.

### 2.2.1 Signal extraction from Nanopore reads

Before extracting signal features from Nanopore reads, two steps are used to process the data: (1) basecall: For *H. sapiens*, *E. coli* R9 2 D reads provided by Simpson *et al.* (2017) and the pUC19 R9 2 D reads provided by Rand *et al.* (2017), we use the basecalling results from the downloaded data. For the R9.4 1 D reads of *H. sapiens* NA12878 and HX1, we use Albacore (version 2.3.1), which is the official basecaller of Oxford Nanopore Technologies, to predict the bases from original signals. (2) re-squiggle: We use the re-squiggle module of nanoraw/tombo (Stoiber *et al.*, 2016) to map the raw electrical signal values to contiguous bases in the genome reference. The nanoraw/tombo corrects the indel errors in Nanopore base call and re-annotates raw signal to match the genomic bases. We use

nanoraw [with BWA-MEM (Li, 2013) for alignment] to re-squiggle all the *H. sapiens*, *E. coli* and pUC19 R9 2 D reads and use tombo [with minimap2 (Li, 2018) for alignment] to re-squiggle R9.4 1 D reads of *H. sapiens* NA12878 and HX1. After the re-squiggle, we extract corresponding raw signals of each matched base in the genome reference. The genome references of *E. coli* and *H. sapiens* used for alignment are *E. coli* K12 MG1655 and GRCh38.p5, respectively. The pUC19 plasmid reference sequence is downloaded from NCBI Nucleotide.

After the re-squiggle, we normalize the raw signals of each read separately by using median shift and median absolute deviation (MAD) scale (Stoiber *et al.*, 2016) as follows:

$$\text{signal}_{\text{norm}} = \frac{\text{signal}_{\text{raw}} - \text{median(signals)}}{\text{MAD(signals)}}, \quad (1)$$

where signals are signal values of a read, $\text{signal}_{\text{raw}}$ and $\text{signal}_{\text{norm}}$ are the raw and the normalized values of a signal, respectively.

For each candidate of methylated site (a C in CpG or CCWGG, an A in GATC), we extract the normalized signals of 17 bp nucleotide sequence centering on it (Supplementary Text and Supplementary Fig. S11A). Then, for each base, we calculate the mean, standard deviation and the number of its signal values. Therefore, for each sample, we have four 17-length vectors as input for following *sequence* module: the 17-mer nucleotides and three other vectors that contain the mean, the standard deviation and the number of signal values of each base, respectively. Meanwhile, we select 360 normalized signal values in the middle of the 17-mer's signal values as input to *signal* feature module, which covers most of signals in 17-mers (Supplementary Fig. S2). If the number of signals in a 17-mer is less than 360, we paddle with zeros.

### 2.2.2 *Sequence* feature module

In *sequence* feature module, we train a BRNN to construct features from four 17-length signal information vectors. Each BRNN (Schuster and Paliwal, 1997) includes a forward RNN and a backward RNN to catch both past and future context. A RNN scans the sequence of data and encodes the sequential information into hidden state vector $h$. We use the long short-term memory (LSTM) RNN (Sak *et al.*, 2014) in our BRNN. Let $x_1, x_2, x_T$ be a sequence of signal features (each time step $x_t$ contains four features: the nucleotide type, the mean value and standard deviation value of signal values that are mapped to current nucleotide, and the number of signal values at current time step). A LSTM RNN will recursively calculate the hidden layer $h$ as follows:

$$i_t = \text{sigmoid}(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci} \odot c_{t-1} + b_i), \quad (2)$$

$$f_t = \text{sigmoid}(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf} \odot c_{t-1} + b_f), \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc} \odot c_t + b_c), \quad (4)$$

$$o_t = \text{sigmoid}(W_{xo}x_t + W_{ho}h_{t-1} + W_{co} \odot c_t + b_o), \quad (5)$$

$$h_t = o_t \odot \tanh(c_t), \quad (6)$$

where $W$ and $b$ are weights and biases in the model; $x$ is the input vector; $i$ is the activation vector of input gate; $f$ is the activation vector of forget gate; $c$ is the cell state vector; $o$ is the activation vector of output gate and $h$ is the output vector of the LSTM hidden unit. Current output $h_t$ of LSTM hidden unit depends on the input $x_t$, previous state $h_{t-1}$, and previous information stored in cell. Then, the output of forward and backward LSTM RNN is combined

$$z_t = h_{t,F} \oplus h_{t,B}. \quad (7)$$

The *sequence* feature module has three forward and three backward layers (Supplementary Text and Supplementary Fig. S11B). For each sample, the *sequence* feature module generates a representation vector $z$, which is inputted to classification module.

### 2.2.3 *Signal* feature module

In *signal* feature module, we train a deep CNN from 360 raw signal values of each targeted base and its close neighbor bases. Our deep CNN is a variation of GoogLeNet (Szegedy *et al.*, 2015) with modified inception block. There is a 1 × 1 conv, a 1 × 3 conv, a 1 × 5 conv, a residual 1 × 3 conv and a 1 × 3 maxpool in our inception block. There are totally 11 inception blocks stacked in the *signal* feature module (Supplementary Text and Supplementary Fig. S11C). We also use batch normalization and ReLU activation between some layers. When doing the operation convolution, the vector $x$ will be calculated as follows:

$$x_c = \text{Conv}(x) = \sum_{i=0}^{l} \sum_{j=0}^{k-1} w_j * x_{i+j}, \quad (8)$$

where $w$ is the weight vector of the convolution kernel; $l$ is the sequence length; $k$ is the 1-dimension convolution kernel size. Then, we apply batch normalization, which can keep the mean and variance fixed without saturation. The batch normalization over a batch data is

$$x_{\text{bn}} = \text{BN}_{\gamma,\beta}(x_c), \quad (9)$$

$$\text{BN}_{\gamma,\beta}(x_c) = \frac{\gamma}{\sqrt{\text{Var}[x_c] + \varepsilon}} \cdot x_c + \left(\beta - \frac{\gamma E[x_c]}{\sqrt{\text{Var}[x_c] + \varepsilon}}\right), \quad (10)$$

where $\gamma$, $\beta$, $\varepsilon$ are the parameters to be learned. Batch normalization can improve the performance and stability of neural networks (Ioffe and Szegedy, 2015).

After batch normalization, we use ReLu as the activation function

$$x_a = \text{Relu}(x_{\text{bn}}) = \begin{cases} 0, \text{ if } x_{\text{bn}} < 0 \\ x_{\text{bn}}, \text{ otherwise} \end{cases}. \quad (11)$$

Finally, a pooling operation is used to summarize the activations of adjacent $k$ neurons by taking their maximum value

$$x_p = p_i(x_a) = \max([x_{a,\ldots i}, \ldots x_{a,\ldots i+k}]). \quad (12)$$

We sequentially select $k$ neurons without overlapping for pooling.

### 2.2.4 Classification module

A fully connected neural network with two hidden layers does the high-level reasoning in classification module. A sigmoid activation function is used in output layer to calculate final output $\hat{y} \in [0, 1]$ of the target site

$$\hat{y} = \frac{1}{1 + e^{-x}}. \quad (13)$$

### 2.2.5 Model training

Model parameters are learned on the training set by minimizing the cross-entropy loss function $L$ as follows:

$$L = z * -\log(\hat{y}) + (1 - z) * -\log(1 - \hat{y}), \quad (14)$$

where $z$ is the true label distribution of training data. We use a batch size of 512 and an initial learning rate 0.001. The learning rate is

adapted by Adam (Kingma and Ba, 2014) and decayed by a factor of 0.1 after every two epochs. We use two strategies to prevent overtraining. First, we use dropout (Srivastava *et al.*, 2014) to regularize the network. We keep the node with probability value of 0.5 at each dropout layer. Second, we use early stopping (Prechelt, 1998). The selection of model is based on its performance on validation dataset. We save the model with current best performance before further training. If the performance of current model is lower than that of saved best model, we don't save the model parameters. The training is stopped if the best performance of the current epoch decreases compared to the best performance of previous epoch.

We choose the *k*-mer window size (17), the number of neural network layers (three for BRNN and 11 for inception layers) of our model and the number of samples for training based on the experiments of DNA CpG methylation detection with *E. coli* R9 2D data (Supplementary Text and Supplementary Fig. S11). The whole module is implemented in Python using tensorflow 1.8.

#### 2.2.6 Data partitioning for experiments

To compare DeepSignal with two HMM based methods: nanopolish (Simpson *et al.*, 2017) and signalAlign (Rand *et al.*, 2017), we first split the whole data of each experiment to training dataset and testing dataset. When training DeepSignal, we further split the training dataset into training and validating datasets. We treat each targeted DNA base at read level as a sample. Numbers of samples for training, validating and testing DeepSignal in each experiment are provided in Supplementary Table S2.

#### 2.2.7 Performance evaluation

We evaluate DeepSignal at both read level and genome level. For each targeted DNA base in a read, DeepSignal outputs two probabilities $P_+$ and $P_-$ [Equation (15)], which represent the probabilities of methylated and unmethylated state, respectively,

$$P_+ = \frac{\hat{y}_+}{\hat{y}_+ + \hat{y}_-}, P_- = \frac{\hat{y}_-}{\hat{y}_+ + \hat{y}_-}, \tag{15}$$

where $\hat{y}_+$ and $\hat{y}_-$ are outputs of sigmoid function calculated by Equation (13), for methylated and unmethylated state, respectively. If $P_+ > P_-$, the targeted DNA base is predicted as methylated, otherwise is predicted as unmethylated. Based on the predicted results, we calculate accuracy, sensitivity, specificity, precision and Area Under receiver operating characteristic Curve (AUC) value of DeepSignal at read level.

To evaluate DeepSignal at genome level, we group the predictions of targeted DNA bases in reads that are aligned to the same position in reference genome. Then, we calculate the methylated probability $P_=$' and the unmethylated probability $P_-$' of each tested site in reference as follows:

$$P'_+ = \frac{1}{n}\sum_{i=1}^{n} P_{+,i}, P'_- = \frac{1}{n}\sum_{i=1}^{n} P_{-,i} \tag{16}$$

where $n$ is the number of targeted DNA bases in reads which are aligned to the site in reference genome, $P_{+,i}$ and $P_{-,i}$ are the methylated probability and the unmethylated probability of the targeted DNA base in $i$th read.

#### 2.2.8 Computational setup

There are three steps of DeepSignal: feature extraction, training and testing. We perform the feature extraction step on a server with 48 CPU processors (Intel(R) Xeon(R) Gold 6126 CPU @ 2.60GHz). And we performed the training and testing steps on a server with 4*12GB TITAN V GPUs. DeepSignal also supports training and testing on CPU severs. The running time and peak memory of DeepSignal on all the tested data in this paper are provided in Supplementary Text and Supplementary Table S3.

## 3 Results

### 3.1 DNA methylation signals in Nanopore reads

Previous studies (Stoiber *et al.*, 2016) already showed that it is possible to distinguish the raw electrical signals of methylated site from those of unmethylated site using statistical tests. Supplementary Figure S3 shows the boxplots of signals in methylated and unmethylated CpG of *E. coli* and CCWGG of *E. coli*. The difference between raw signals of targeted base C (the cytosine in CpG and the inner cytosine of CCWGG) is not significant between methylated and unmethylated reads. Meanwhile, the raw signal distribution of bases around targeted base C show significant difference between methylated and unmethylated reads. For example, the signals of base T and G in CCWGG have significant difference between methylated and unmethylated reads, while the signals of target C base show no difference (Supplementary Fig. S3B). Although statistics based methods can identify signal difference of bases around targeted base, they ignored the relationship between bases. It is important to catch both signal and sequence information of bases around the targeted base for DNA methylation state prediction.

### 3.2 Evaluation of DeepSignal on prediction of CpG methylation (5mC) data of *E. coli* and *H. sapiens*

To evaluate the performance of DeepSignal, we first test it using DNA CpG methylation (5mC) data of *E. coli* and *H. sapiens* (Simpson *et al.*, 2017). Both data include Nanopore reads of PCR-amplified DNA that was either untreated or treated with CpG methyltransferase M.SssI. In the M.SssI-treated data, the M.SssI methylase converts nearly all of the unmethylated CpGs to 5mCpGs (Simpson *et al.*, 2017). As in Simpson *et al.* (2017), all PASS and FAIL reads with high mapping quality to reference genome (quality score greater than 20) are used to extract signals by nanoraw (Stoiber *et al.*, 2016). We only select CpGs in reads which have signals in both template and complement strands. There are 49 378 592 CpGs in the reads of *E. coli* data and 3 120 754 CpGs in the reads of *H. sapiens* data. We first perform validation in each genome. We randomly select 50% of the *H. sapiens* and *E. coli* reads to train models, and test on the remaining 50% of data. To validate the robustness of DeepSignal, we randomly repeat train-test split ten times. Then, we perform a cross genome validation. We train DeepSignal using CpG sites of *E. coli* genome and test it on CpG sites of *H. sapiens* genome. We train and test nanopolish using the same data.

Both DeepSignal and nanopolish learn two models from template and complement data separately. During testing, we predict the methylation site on template and complement separately using different models. As in nanopolish, assuming that there are no hemi-methylated CpGs, we only output the final prediction on the sites aligned to the forward strand of the genome reference, which is summarized by predicted probabilities of methylation sites from both strands of 2D nanopore reads. The performances of both tools are summarized at read level. All singleton CpGs in reads, of which there are no other CpGs in the up and down 10 bp region (Simpson *et al.*, 2017), are evaluated. As shown in Table 2, for validation in each genome, DeepSignal achieves higher performance on all

**Table 2.** Performance comparison of DeepSignal and nanopolish on predicting methylation states of DNA CpGs at read level using *H. sapiens* and *E. coli* R9 2D reads

| Train/test | Method | Accuracy | Sensitivity | Specificity | Precision | AUC |
|---|---|---|---|---|---|---|
| *E. coli*/*H. sapiens* | DeepSignal | 0.938 | 0.944 | 0.930 | 0.935 | 0.983 |
| | nanopolish | 0.894 | 0.844 | 0.947 | 0.944 | 0.965 |
| *E. coli*/*E. coli* | DeepSignal | 0.957 (±0.006) | 0.965 (±0.003) | 0.952 (±0.013) | 0.938 (±0.016) | 0.993 (±0.002) |
| | nanopolish | 0.886 (±0.001) | 0.870 (±0.001) | 0.899 (±0.000) | 0.866 (±0.001) | 0.956 (±0.000) |
| *H. sapiens*/*H. sapiens* | DeepSignal | 0.947 (±0.001) | 0.953 (±0.003) | 0.942 (±0.005) | 0.946 (±0.004) | 0.988 (±0.000) |
| | nanopolish | 0.908 (±0.000) | 0.893 (±0.000) | 0.925 (±0.000) | 0.927 (±0.001) | 0.968 (±0.000) |

*Note*: For the experiments of *E. coli*/*E. coli* and *H. sapiens*/*H. sapiens*, values are average and standard deviation of 10 replicated train-test splits.

evaluation criteria: accuracy, sensitivity, specificity, precision and AUC, although DeepSignal only has a slightly higher specificity and precision comparing to nanopolish in *H. sapiens* (0.942 versus 0.925 for specificity and 0.946 versus 0.927 for precision). For the evaluation cross genomes, comparing to nanopolish, DeepSignal has substantially higher performance on accuracy (0.938 versus 0.894), sensitivity (0.944 versus 0.844) and AUC (0.983 versus 0.965), while has a slightly less performance on specificity (0.930 versus 0.947) and precision (0.935 versus 0.944). This result shows that DeepSignal is more powerful than nanopolish in predicting the methylation states of CpG sites.

### 3.3 Evaluation of DeepSignal on G*A*TC and C*C*WGG methylation data of pUC19 plasmid DNA

We then test DeepSignal using G*A*TC (6 mA) and C*C*WGG (5mC) methylation data, which contains Nanopore reads of *dam* and *dcm* methyltransferase treated pUC19 vector DNA and its PCR-amplified controls (Rand *et al.*, 2017). According to Rand *et al.* (2017), in the sequenced pUC19 vector DNA, methyltransferase *dam* and *dcm* methylate all adenines in G*A*TC motifs and the inner cytosines in C*C*WGG motifs to 6 mA and 5mC, respectively. The pUC19 plasmid reference is 2686 bp long and contains 30 G*A*TC motifs and 10 C*C*WGG motifs. To evaluate DeepSignal, we select the reads of which the 2D sequences cover more than 2600 bp of the pUC19 DNA reference. After the alignment using BWA-MEM (Li, 2013), 12 844 PASS reads of methylated genome (gDNA) and 18 892 PASS reads of unmethylated genome (pcrDNA) are kept. For both G*A*TC and C*C*WGG methylation detections, we use 40% of the reads in pUC19 data to train the DeepSignal models. Then, we randomly select 40 methylated and 40 unmethylated reads from the remaining 60% for testing and repeat 100 times. We further repeat the train-test split ten times to validate the robustness of DeepSignal. We use the same reads to train and test DeepSignal and signalAlign.

We compare the performance of DeepSignal with signalAlign (Rand *et al.*, 2017) at genome level. For each targeted site in reference genome, its methylation state is summarized from the predicted probabilities of all reads aligned to the site. We train the model using the template strand only. For predicting methylation states of 6 mA in G*A*TC motif, DeepSignal achieves 0.999 accuracy while signalAlign just gets accuracy of 0.908 (Table 3). Moreover, DeepSignal achieves much higher specificity (1.000 versus 0.846) and precision (1.000 versus 0.865). For predicting methylation states of 5mC in C*C*WGG motif, DeepSignal achieves accuracy of 0.997 while signalAlign gets accuracy of 0.962 (Table 3). Similar to the performance on G*A*TC methylation data, DeepSignal achieves similar sensitivity, but higher specificity and precision.

### 3.4 Comparison with bisulfite sequencing on detecting DNA CpG methylation states in *H. sapiens*

The bisulfite sequencing method is the current 'gold standard' for 5mC modification detection in DNA. To further evaluate the effectiveness of DeepSignal for predicting DNA methylation state, we compare it with the results of bisulfite sequencing of two *H. sapiens* samples: HX1 and NA12878 (Consortium *et al.*, 2012). Previous study (Jain *et al.*, 2018) already showed that nanopolish can achieve high correlation with bisulfite sequencing with high read coverage.

To train DeepSignal of *H. sapiens* CpG using Nanopore reads, we first select high-confidence methylated and unmethylated DNA CpG sites based on bisulfite sequencing data (see Section 2). We train DeepSignal using 10 million reads mapped to high-confidence methylated CpGs and 10 million reads mapped to high-confidence unmethylated CpGs, which are selected from 10× reads of HX1 native DNA and 5× HX1 WGA reads. Then, we test DeepSignal on remaining 40× HX1 Nanopore reads and 30× NA12878 reads. During the training and testing, all PASS Nanopore reads are used to extract signals. We run nanopolish (v0.10.1) with built-in models on the same test data for comparison.

We first evaluate DeepSignal and nanopolish for detecting methylation states of selected high-confidence DNA CpGs at read level. We first use them to predict methylation states of singleton CpGs (Simpson *et al.*, 2017) in read. As shown in Table 4, DeepSignal achieves substantially higher performance for single CpG prediction on accuracy, specificity, sensitivity precision and AUC, comparing to those of nanopolish. Previously, nanopolish reported low performance on predicting methylation states of CpGs mixed with methylated and unmethylated CpGs appearing in a close neighborhood (10 bp) (Simpson *et al.*, 2017). We further test whether DeepSignal can predict methylation states of CpGs mixed with methylated and unmethylated. We select a set of high-confidence methylated or unmethylated CpGs with at least another high-confidence CpG with different methylation state in up or down 10 bp region. We have obtained 11 127 methylated CpGs and 11 711 unmethylated CpGs from HX1, as well as 5654 methylated and 5690 unmethylated CpGs from NA12878. We evaluate DeepSignal and nanopolish using the reads aligned to selected methylated and unmethylated sites. As shown in Table 4, DeepSignal can still achieve 0.866 and 0.84 accuracy on mixed CpG sites of HX1 and NA12878, respectively, while nanopolish can only have 0.526 and 0.535 accuracy, respectively. Moreover, DeepSignal has 0.943 and 0.917 AUC values for testing on two samples, while nanopolish only has 0.545 and 0.554 AUC values, respectively. The performances of DeepSignal on mixed CpGs are only slightly less than those on singleton CpGs. These results demonstrate that DeepSignal can catch the methylation characteristic of CpGs better than nanopolish does.

**Table 3.** Performance comparison of DeepSignal and signalAlign on predicting methylation states of 6 mA in GATC motif and 5mC in CCWGG motif of pUC19 DNA using template reads only

| Method | 6mA in GATC motif | | | | | 5mC in CCWGG motif | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Sensitivity | Specificity | Precision | AUC | Accuracy | Sensitivity | Specificity | Precision | AUC |
| DeepSignal | 0.999 (±0.001) | 0.998 (±0.002) | 1.000 (±0.000) | 1.000 (±0.000) | 1.000 (±0.000) | 0.997 (±0.002) | 0.994 (±0.005) | 1.000 (±0.000) | 1.000 (±0.000) | 1.000 (±0.000) |
| signalAlign | 0.908 (±0.015) | 0.970 (±0.007) | 0.846 (±0.031) | 0.865 (±0.024) | 0.960 (±0.006) | 0.962 (±0.018) | 0.996 (±0.005) | 0.929 (±0.035) | 0.938 (±0.030) | 0.959 (±0.007) |

*Note*: To calculate the values of accuracy metrics, we first calculate the average values of 100 replicated tests in each train-test split. Then, we calculate the average and standard deviation values of 10 train-test splits.

**Table 4.** Performance comparison of DeepSignal and nanopolish on predicting methylation states of CpGs at read level using *H. sapiens* R9.4 1D data

| Data | Method | Accuracy | Sensitivity | Specificity | Precision | AUC |
|---|---|---|---|---|---|---|
| Singleton CpG prediction | | | | | | |
| HX1 | DeepSignal | 0.894 | 0.893 | 0.923 | 0.995 | 0.963 |
| | nanopolish | 0.823 | 0.823 | 0.817 | 0.986 | 0.897 |
| NA12878 | DeepSignal | 0.900 | 0.881 | 0.918 | 0.902 | 0.955 |
| | nanopolish | 0.835 | 0.829 | 0.841 | 0.817 | 0.907 |
| Mixed CpG prediction | | | | | | |
| HX1 | DeepSignal | 0.866 | 0.819 | 0.910 | 0.895 | 0.943 |
| | nanopolish | 0.526 | 0.595 | 0.462 | 0.509 | 0.545 |
| NA12878 | DeepSignal | 0.840 | 0.793 | 0.887 | 0.874 | 0.917 |
| | nanopolish | 0.535 | 0.614 | 0.457 | 0.529 | 0.554 |

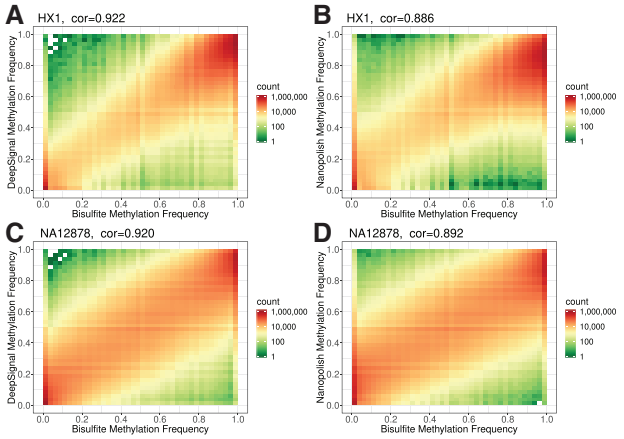


**Fig. 2.** Comparison of methylation frequencies of CpGs calculated by DeepSignal/nanopolish with those from bisulfite sequencing (cor is Pearson correlation). (**A**, **B**) HX1; (**C**, **D**) NA12878

We then evaluate the correlation between DeepSignal and bisulfite sequencing. We calculate the methylation frequencies for all DNA CpGs at genome level in HX1 and NA12878 based on DeepSignal predictions of Nanopore reads. We combine reads mapped to both forward and backward strand of the *H. sapiens* genome to calculate methylation frequency of each CpG. We select CpGs that are covered with at least five reads for evaluation. We calculate the Pearson correlation between methylation frequencies calculated from bisulfite sequencing and those calculated using DeepSignal. Only CpGs whose methylation frequencies are calculated by both bisulfite sequencing and DeepSignal are used for calculating correlation. For both HX1 and NA12878, DeepSignal shows a better correlation with bisulfite sequencing than nanopolish does (Fig. 2). DeepSignal has correlation values of 0.923 and 0.920 with the bisulfite sequencing results for HX1 and NA12878, respectively, while nanopolish only has 0.886 and 0.892 for HX1 and NA12878, respectively. We then further evaluate the methylation frequency correlations in CpG islands of *H. sapiens* genome. As shown in Supplementary Figure S4, DeepSignal still gets a better consistency with bisulfite sequencing (higher Pearson correlations) in CpG islands than nanopolish does.

It is important to detect both highly methylated and lowly methylated CpGs in the genome. We then evaluate the overlapping of highly methylated (methylation frequency >= 0.7) and lowly methylated (methylation frequency <= 0.3) CpGs detected by DeepSignal with those detected by bisulfite sequencing. As shown in

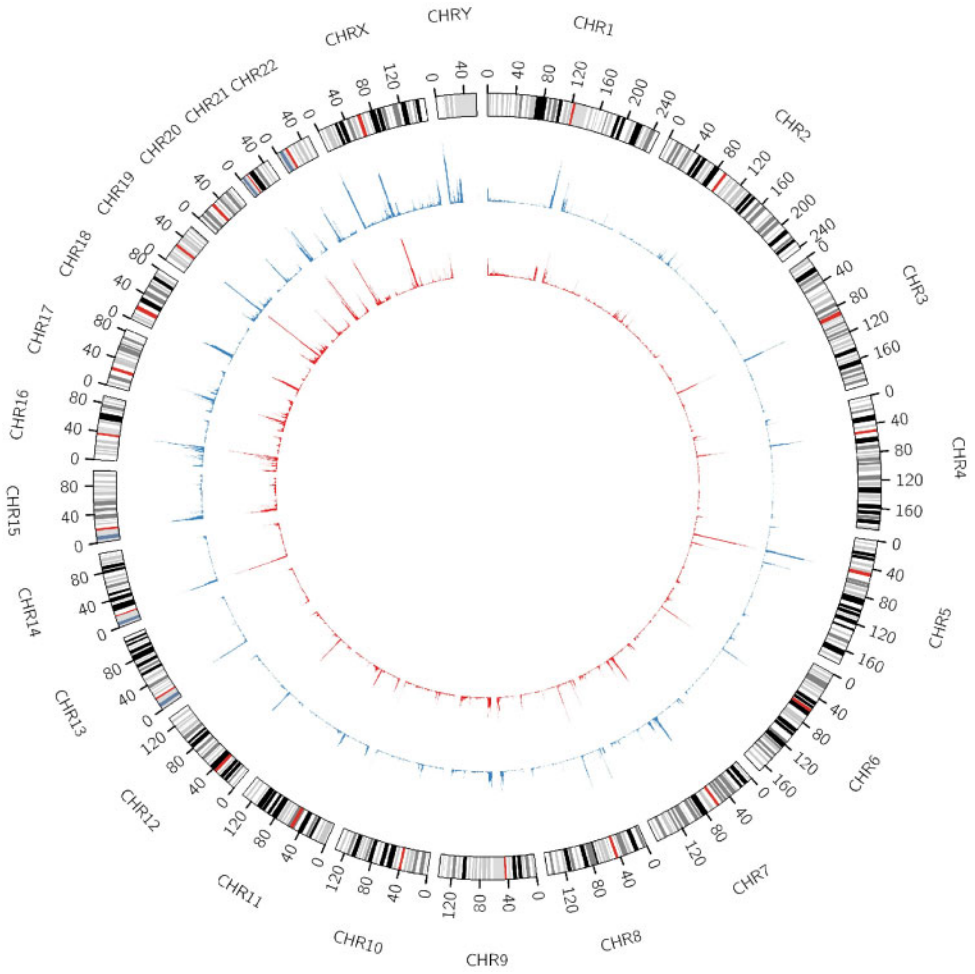


**Fig. 3.** Distribution of number of CpGs whose methylation states can be detected by DeepSignal only. Inner blue cycle: HX1; inner red cycle: NA12878; outer cycle: reference (the chromosomes are binned into 1 000 000-bp windows. The centromeric region is indicted by red bar in each chromosome)

Supplementary Figure S5, there is 74.7% overlapping between lowly methylated CpGs and 86.7% overlapping between highly methylated CpGs detected in HX1 by DeepSignal and bisulfite sequencing. The overlap is 70% for lowly methylated CpGs and 73.8% for highly methylated CpGs in NA12878. The difference of overlapping rates between HX1 and NA12878 may be due to the different coverage of Nanopore reads. Those results imply that DeepSignal can achieve high performance in detection of both highly methylated and lowly methylated CpGs.

Due to drawbacks of second generation sequencing technology, such as GC bias and short reads, bisulfite sequencing usually is not able to detect methylation states of CpGs in high GC, highly repeat and transposable regions in genome. On the other hand, Nanopore sequencing does not have GC bias and has much longer read length. We then evaluate the coverage of CpGs detected by DeepSignal. There are 29 401 360 CpGs in the 24 chromosomes of *H. sapiens* GRCh38.p5 genome reference. With at least 5× coverage, for HX1 genome, DeepSignal can predict methylation states of 1 668 648 or 5.68% more CpGs (Supplementary Fig. S6A). For NA12878 genome, DeepSignal can predict methylation states of 1 868 545 or 6.36% more CpGs (Supplementary Fig. S6B). We have then examined seven specific regions in *H. sapiens* GRCh38/hg38 genome (Casper *et al.*, 2017): Centromere, CpG islands, Repeats, Exon, Intron, 5' UTR and 3' UTR (Supplementary Fig. S7), which are enriched with CpGs predicted by DeepSignal only (Fig. 3 and Supplementary Fig. S7). A significant amount of CpGs that can only be detected by DeepSignal are at centromeric regions: 28.3 and 25.6% for HX1 and NA12878, respectively. It has been reported that DNA methylation plays an important role in centromere identity (Zhang *et al.*, 2008) and centromere evolution (Ichikawa *et al.*, 2017). Thus, with more methylated sites detected at centromere regions, DeepSignal may help to further understand the functionality of centromeric region. There are also over 18% of those CpGs in CpG islands and Intron regions for both HX1 and NA12878. As shown in Supplementary Figure S8, among those sites which can only be detected by DeepSignal, there are 45.7% highly methylated CpGs and 26.7% lowly methylated CpGs in HX1. In NA12878, there are 22.8% highly methylated CpGs and 46.8% lowly methylated CpGs. There are also 86 338 CpGs in HX1 and 81 507 CpGs in NA12878 that cannot be detected by DeepSignal. Over 50% of those CpGs in both HX1 and NA12878 are in centromeric regions (Supplementary Fig. S7B). There are less coverage of reads in those regions because DeepSignal only picks the best match in alignment step.

### 3.5 Evaluation of the data coverage effect on DeepSignal

The coverage of reads usually affects the DNA methylation state prediction. We further evaluate DeepSignal at genome level with

**Table 5.** Accuracies of DeepSignal and nanopolish for predicting methylation states of DNA CpGs at genome level under different coverages of *H. sapiens* R9.4 1D data (left) and Pearson correlations of methylation frequencies calculated by DeepSignal/nanopolish and bisulfite sequencing of under different coverages (right)

| Data | Method | CpG site prediction | | | | | | The Pearson correlation | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1× | 2× | 5× | 10× | 20× | 30× | 5× | 10× | 20× | 30× |
| HX1 | DeepSignal | 0.924 | 0.948 | 0.983 | 0.995 | 0.998 | 0.999 | 0.845 | 0.875 | 0.903 | 0.915 |
| | nanopolish | 0.851 | 0.880 | 0.940 | 0.976 | 0.992 | 0.995 | 0.753 | 0.780 | 0.851 | 0.873 |
| NA12878 | DeepSignal | 0.923 | 0.943 | 0.979 | 0.993 | 0.997 | 0.998 | 0.854 | 0.879 | 0.909 | 0.920 |
| | nanopolish | 0.863 | 0.891 | 0.946 | 0.979 | 0.992 | 0.995 | 0.800 | 0.833 | 0.875 | 0.892 |

*Note*: For each coverage, the reads are randomly shuffled and selected. Values in the table are average of ten replicated tests.

sampled sub-set of reads. For CCWGG and GATC methylation data of pUC19 plasmid, we evaluate the coverage effect on DeepSignal by sampling different number of reads. As shown in Supplementary Table S1, when the model trained from template strand data is used for predicting GATC methylation and CCWGG methylation in pUC19, DeepSignal can achieve above 0.9 accuracy with only two sampled reads, while signalAlign needs 30 or more sampled reads to have above 0.9 accuracy. DeepSignal also achieves higher sensitivities, specificities, precision and AUC than signalAlign in predicting GATC methylation (Supplementary Fig. S9A–D). For predicting CCWGG methylation, DeepSignal achieves similar sensitivities and higher specificities, precision and AUC than signalAlign (Supplementary Fig. S9E–H). The results demonstrate that DeepSignal can achieve high prediction performance with much less read coverage than those needed by signalAlign.

Then, for HX1 and NA12878 data, we evaluate DeepSignal with different coverage of reads. We first assess the accuracy of methylation state prediction of high-confidence DNA CpGs. As shown in Table 5, DeepSignal can achieve 0.924 and 0.923 accuracy with only 1× coverage of reads for HX1 and NA12878, respectively. For very low coverage (1× and 2×), DeepSignal achieves significant higher accuracies than nanopolish does (Table 5). For more than 5× coverage, nanopolish needs double coverage of reads to achieve the same accuracy of DeepSignal. As shown in Supplementary Fig. S10A–D, at low coverage (1× and 2×) of HX1 data, DeepSignal achieves slightly higher precision but much higher sensitivities, specificities and AUC than nanopolish. DeepSignal also achieves much higher sensitivities, specificities, precision and AUC than nanopolish achieves for NA12878 at low coverage (Supplementary Fig. S10E–H). Second, we evaluate the correlation between methylation frequencies calculated from bisulfite sequencing and those calculated using DeepSignal with different coverages of reads. Table 5 shows that DeepSignal can achieve above 0.9 Pearson correlations with bisulfite sequencing for both HX1 and NA12878 with only 20× coverage of reads, while nanopolish cannot achieve above 0.9 even with 40× coverage for HX1. With only 5× coverage of reads, DeepSignal has achieved 0.845 and 0.854 Pearson correlations with bisulfite sequencing for HX1 and NA12878, respectively, which nanopolish only has achieved 0.753 and 0.800 correlations, respectively. DeepSignal can achieve high performance on CpG methylation state prediction with much less read coverage than those needed by nanopolish.

## 4 Discussion

In this study, we develop DeepSignal, a deep learning method to detect DNA methylation states from Nanopore sequencing reads. Testing DeepSignal on Nanopore reads of pUC19, *E. coli* and *H. sapiens*, we show that it can achieve higher accuracy on detect both DNA 6 mA and 5mC sites comparing to previous HMM based methods. DeepSignal can detect DNA 5mC in both CpGs and CCWGG motifs. Moreover, DeepSignal can predict methylation states of both single CpGs and mixed CpGs. Furthermore, DeepSignal achieves similar performance on different methylation bases and different methylation motifs, while other methods, like signalAlign, have higher performance on 5mC methylation site than on 6 mA methylation site. DeepSignal is able to catch the essential characteristic of DNA methylation and is more robust to detect different DNA methylation types.

Due to still relative high cost of Nanopore sequencing, it is useful to identify methylation sites with low coverage of reads. We demonstrate that DeepSignal can achieve 90% above accuracy for detecting DNA 5mC and 6 mA using only 2× coverage of reads. DeepSignal requires much lower coverage than those required by HMM and statistics based methods. Furthermore, experiments on CpG methylation state prediction in *H. sapiens* HX1 and NA12878 have shown that DeepSignal achieves higher correlation with bisulfite sequencing than HMM based methods do. To obtain 0.9 above correlation with bisulfite sequencing, DeepSignal only needs 20× coverage of reads. Furthermore, compared to bisulfite sequencing, DeepSignal can predict methylation states of 5% more CpGs that previously cannot be predict by bisulfite sequencing. Currently, DeepSignal can be used immediately for predicting methylation states of DNA CpGs in eukaryotes and prokaryotes. With more training data of other types of base modification available in the future, we will train DeepSignal to detect more types of base modification. DeepSignal may become a new method for detecting methylation states of DNA bases in the future, which will help improving the understanding of epigenetic mechanisms of different biological processes, such as aging, cancer and mental development.

## Acknowledgements

We acknowledge the ENCODE Consortium and the ENCODE production laboratory(s) generating the ENCFF835NTC dataset. We are grateful to Miten Jain and Arthur C. Rand for providing their methylation data.

## Funding

This work was supported in part by the National Natural Science Foundation of China under Grants (Nos. 61420106009, 61732009, 61828205 and 61772557), 111 Project (No. B18059), Hunan Provincial Science and Technology Program (No. 2018wk4001) to J.W., the U. S. National Institute of Food and Agriculture (NIFA) under grant 2017-70016-26051 and the U.S. National Science Foundation (NSF) under grants ABI-1759856 to F.L.

*Conflict of Interest*: none declared.

## References


Bergman,Y. and Cedar,H. (2013) Dna methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.*, **20**, 274.

Casper,J. *et al.* (2017) The UCSC genome browser database: 2018 update. *Nucleic Acids Res.*, **46**, D762–D769.

Consortium,E.P. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57.

Davis,B.M. *et al.* (2013) Entering the era of bacterial epigenomics with single molecule real time dna sequencing. *Curr. Opin. Microbiol.*, **16**, 192–198.

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

Gonzalo,S. (2010) Epigenetic alterations in aging. *J. Appl. Physiol.*, **109**, 586–597.

Ichikawa,K. *et al.* (2017) Centromere evolution and CPG methylation during vertebrate speciation. *Nat. Commun.*, **8**, 1833.

Ioffe,S. and Szegedy,C. (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *Proceedings of the 32nd International Conference on Machine Learning*. pp. 448–456. http://proceedings.mlr.press/v37/ioffe15.html.

Jain,M. *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, **36**, 338.

Kingma,D.P. and Ba,J. (2014) Adam: a method for stochastic optimization. *arXiv*:1412.6980.

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

Laszlo,A.H. *et al.* (2013) Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proc. Natl. Acad. Sci. USA*, **110**, 18904–18909.

Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*:1303.3997.

Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **1**, 7.

Liu,Q. *et al.* (2019) Nanomod: a computational tool to detect DNA modifications using nanopore long-read sequencing data. *BMC Genomics*, **20**, 78.

Lund,G. *et al.* (2004) DNA methylation polymorphisms precede any histological sign of atherosclerosis in mice lacking apolipoprotein e. *J. Biol. Chem.*, **279**, 29147–29154.

McIntyre,A.B. *et al.* (2017) *Nanopore Detection of Bacterial DNA base Modifications*. Cold Spring Harbor Laboratory, doi: 10.1101/127100.

Prechelt,L. (1998) *Early Stopping—but When? In Neural Networks: Tricks of the Trade*. Springer Berlin Heidelberg, pp. 55–69.

Rand,A.C. *et al.* (2017) Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods*, **14**, 411.

Sak,H. *et al.* (2014) Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv*:1402.1128.

Schatz,M.C. (2017) Nanopore sequencing meets epigenetics. *Nat. Methods*, **14**, 347.

Schreiber,J. *et al.* (2013) Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proc. Natl. Acad. Sci.*, **110**, 18910–18915.

Schübeler,D. (2015) Function and information content of dna methylation. *Nature*, **517**, 321.

Schuster,M. and Paliwal,K.K. (1997) Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, **45**, 2673–2681.

Simpson,J.T. *et al.* (2017) Detecting dna cytosine methylation using nanopore sequencing. *Nat. Methods*, **14**, 407.

Smith,Z.D. and Meissner,A. (2013) Dna methylation: roles in mammalian development. *Nat. Rev. Genet.*, **14**, 204.

Srivastava,N. *et al.* (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.

Stoiber,M.H. *et al.* (2016) De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. Cold Spring Harbor Laboratory, doi:10.1101/094672.

Szegedy,C. *et al.* (2015) Going deeper with convolutions. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9.

Xiao,C.-L. *et al.* (2018) N6-methyladenine DNA modification in the human genome. *Mol. Cell*, **71**, 306–318.

Yao,B. *et al.* (2018) Active $N^6$-methyladenine demethylation by DMAD regulates gene expression by coordinating with polycomb protein in neurons. *Mol. Cell*, **71**, 848–857.

Yue,Y. *et al.* (2015) RNA N6-methyladenosine methylation in post-transcriptional gene expression regulation. *Genes Dev.*, **29**, 1343–1355.

Zhang,W. *et al.* (2008) Epigenetic modification of centromeric chromatin: hypomethylation of DNA sequences in the CENH3-associated chromatin in Arabidopsis thaliana and maize. *Plant Cell*, **20**, 25–34.

Zhu,S. *et al.* (2018) Mapping and characterizing N6-methyladenine in eukaryotic genomes using single-molecule real-time sequencing. *Genome research*, **28**, 1067–1078.