Bioinformatics, 38(5), 2022, 1235–1243
https://doi.org/10.1093/bioinformatics/btab745
Advance Access Publication Date: 28 October 2021
Original Paper

Sequence analysis

DeepMP: a deep learning tool to detect DNA base modifications on Nanopore sequencing data

Jose Bonet [1,2,*,†], Mandi Chen [3,4,*,†], Marc Dabad[5,6], Simon Heath[5,6], Abel Gonzalez-Perez [1,2], Nuria Lopez-Bigas[1,2,7] and Jens Lagergren[3,4]

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08028 Barcelona, Spain, [2]Research Program on Biomedical Informatics, Universitat Pompeu Fabra, 08002 Barcelona, Catalonia, Spain, [3]Department of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 114 28 Stockholm, Sweden, [4]Science for Life Laboratory, 171 65 Solna Stockholm, Sweden, [5]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain, [6]Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain and [7]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain

*To whom correspondence should be addressed.
†The authors wish it to be known that these authors contributed equally.
Associate Editor: Pier Luigi Martelli

Received on June 28, 2021; revised on October 20, 2021; editorial decision on October 22, 2021; accepted on October 25, 2021

Abstract

Motivation: DNA methylation plays a key role in a variety of biological processes. Recently, Nanopore long-read sequencing has enabled direct detection of these modifications. As a consequence, a range of computational methods have been developed to exploit Nanopore data for methylation detection. However, current approaches rely on a human-defined threshold to detect the methylation status of a genomic position and are not optimized to detect sites methylated at low frequency. Furthermore, most methods use either the Nanopore signals or the basecalling errors as the model input and do not take advantage of their combination.

Results: Here, we present DeepMP, a convolutional neural network-based model that takes information from Nanopore signals and basecalling errors to detect whether a given motif in a read is methylated or not. Besides, DeepMP introduces a threshold-free position modification calling model sensitive to sites methylated at low frequency across cells. We comprehensively benchmarked DeepMP against state-of-the-art methods on *Escherichia coli*, human and pUC19 datasets. DeepMP outperforms current approaches at read-based and position-based methylation detection across sites methylated at different frequencies in the three datasets.

Availability and implementation: DeepMP is implemented and freely available under MIT license at https://github.com/pepebonet/DeepMP.

Contact: jose.bonet@irbbarcelona.org or mandiche@kth.se

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

1 Introduction

Chemical modifications of nucleotides are important epigenetic markers. These DNA modifications play a crucial role in regulating the expression of genes, and cellular responses to stimuli (Bergman and Cedar, 2013; Jones, 2012; Schübeler, 2015). Among all the modifications, one of the most prevalent and widely studied is DNA methylation, in particular at cytosines. Its relevance to fundamental biological processes such as embryonic development (Lund *et al.*, 2004), aging (Gonzalo, 2010) and diseases (Grønbaek *et al.*, 2007) has highlighted the importance of accurate genome-wide profiling techniques.

Both short- and long-read sequencing technologies are exploited to identify methylated genomic sites. These approaches are associated with different experimental and computational limitations. For instance, the Bisulfite sequencing (Miura *et al.*, 2012) with short reads suffers from limited conversion efficiency (Cytosine to Uracil). Also, the amplification bias and the uncertainty of mapping large repetitive regions add to the experimental complexity of its use for methylation detection. Long-read-based technologies, such as Nanopore and PacBio sequencing approaches, have appeared more recently, and as a result, computational methods for DNA methylation detection are currently under development. Although PacBio

Single-Molecule Real-Time (SMRT) approach can detect DNA methylation directly (Flusberg *et al.*, 2010), low signal-to-noise ratios and coverage requirements have limited its application (Davis *et al.*, 2013; Zhu *et al.*, 2018). Conversely, Nanopore sequencing overcomes these limitations as no amplification and prior enzymatic or chemical treatment steps are required, thus supporting the analysis of DNA molecules harboring modifications in their native state (Laszlo *et al.*, 2013; Schatz, 2017; Schreiber *et al.*, 2013; Wescoe *et al.*, 2014). Therefore, Nanopore sequencing has been recently established as the state-of-the-art long-read sequencing approach to detect DNA methylation.

In recent years, several models have contributed to the improvement of Nanopore's methylation detection accuracy (Liu *et al.*, 2019a,b; McIntyre *et al.*, 2019; Ni *et al.*, 2019; ONT Megalodon, 2021; Rand *et al.*, 2017; Simpson *et al.*, 2017; Stoiber *et al.*, 2016). The majority of these methods focus on directly analyzing the output signals of the Nanopore device and do not make use of the basecalling errors (Liu *et al.*, 2019b; McIntyre *et al.*, 2019; Ni *et al.*, 2019; Rand *et al.*, 2017; Simpson *et al.*, 2017; Stoiber *et al.*, 2016). These models have shown the ability to accurately call the methylated status for a target motif at the level of individual reads (read-based calling). However, the position-based calling (across all reads) relies on a human-defined threshold. An alternative approach to detect DNA methylations through Nanopore uses basecalling errors as model features (Liu *et al.*, 2019a). Methods based solely on this approach can, thus far, only identify methylated genomic positions, rather than reveal the methylation status of every read covering the site. Therefore, these methods tend to underpredict sites methylated at low frequency across cells, i.e. sites at which most reads covering the base will be unmodified. A recent study (Yuen *et al.*, 2021) benchmarking some of the latest published methods reported that tools such as Nanopolish (Simpson *et al.*, 2017) and Tombo (Stoiber *et al.*, 2016) tend to overestimate the number of methylated sites. In contrast, Guppy (ONT Guppy, 2021) suffers from an overestimation of the number of unmethylated reads. Megalodon (ONT Megalodon, 2021) and DeepSignal (Ni *et al.*, 2019) reported the best results overall.

We hypothesized that the combination of basecalling errors and current signals could improve the detection of methylated cytosines in the DNA over the use of only one of them. Thus, we developed DeepMP, a deep learning model that exploits the errors and signals from the data. DeepMP includes a further innovation, a supervised Bayesian model to call the position-based methylation, which is—to our knowledge—unique. It comprises a statistical approach based on the information from individual reads covering a position that improves the detection of the methylation status of a genomic position. We show that DeepMP outperforms state-of-the-art methods DeepSignal (Ni *et al.*, 2019), Nanopolish (Simpson *et al.*, 2017), Guppy (ONT Guppy, 2021) and Megalodon (ONT Megalodon, 2021) in the tasks of detecting modified reads and positions methylated at varying frequencies across *Escherichia coli*, human and pUC19 Nanopore long-read data.

2 Materials and methods

2.1 Datasets

Three datasets, K12 ER2925 (*E.coli*), NA12878 (human) and pUC19 (plasmid DNA) were useed to evaluate the performance and benchmark DeepMP. Supplementary Table S1 summarizes all datasets and the number of reads available for each sample. The datasets used in this study contain methylation on the cytosines (*E.coli* and human) and adenine (pUC19) residues. DeepMP was trained to distinguish unmethylated cytosines (C) from 5-Methylcytosines (5mC) at CpG motifs in the first two datasets and unmethylated adenines (A) from 6-methyladenines (6 mA) at GATC motifs in the latter.

2.1.1 CpG methylation data

Nanopore reads from *E.coli* (K12 ER2925) (Simpson *et al.*, 2017) were downloaded from the European Nucleotide Archive under accession number PRJEB13021. The dataset contains reads obtained from *E.coli* treated with M.SssI, which lead to methylation of ~95% of the CpGs (5mC methylated), and PCR-amplified (negative control), which are completely unmethylated. Although reads from both Nanopore R7.3 and R9 flow cells are included, R7.3 flow cells do not provide raw signals, and therefore only R9 reads were used. The ground truth for *E.coli* is generated by selecting treated samples as fully methylated and the control as fully unmethylated.

Human Nanopore reads from the NA12878 datasets were downloaded from the European Nucleotide Archive under accession number PRJEB23027 (Jain *et al.*, 2018). Reads are obtained from 5 sequencing studies (Norwich, UCSC, Bham, Notts and UBC). However, only the Norwich subset is used (Supplementary Table S1). Datasets were basecalled by Guppy 4.4.2 [available to members of the Nanopore community at nanoporetech.com (ONT Guppy, 2021)].

2.1.2 Bisulfite sequencing data

Bisulfite sequencing results of the NA12878 datasets were downloaded from ENCODE (ENCFF835NTC) (The ENCODE Project Consortium, 2012). Both replicates available allow the labeling of high-quality methylated and unmethylated positions. We follow the approach of Liu *et al.* (2019b) to characterize the methylation status of each position. If a cytosine in a given position in the reference genome (GRCh38) contained >90% of methylations in both replicates, that position was considered to be modified and hence completely methylated. On the other hand, if a cytosine has 0% methylations in both replicates of bisulfite sequencing, the position was considered unmodified and thus completely unmethylated.

2.1.3 6mA methylation data

Raw Nanopore reads from pUC19 were downloaded from NCBI under accession number SRR5219626. pUC19 plasmids were cloned in *E.coli* grown either in the presence (treated) or absence (untreated) of Dam methyltransferase (Supplementary Table S1). The presence of this enzyme leads to the complete methylation adenines (m6A) in GATC motifs (Rand *et al.*, 2017). Sequences were also basecalled using Guppy 4.4.2 (ONT Guppy, 2021).

2.2 Benchmarking methylation detection with similar tools

We benchmarked DeepMP against four existing tools: Nanopolish (Simpson *et al.*, 2017), Megalodon (ONT Megalodon, 2021), DeepSignal (Ni *et al.*, 2019) and Guppy (ONT Guppy, 2021). The different tools are explained in Supplementary Section S1.1 of Supplementary Methods.

2.3 Preprocessing and feature extraction

DeepMP consists of two different modules: a sequence module for handling the signals (Nanopore currents) and another module to process basecalling errors (Fig. 1A and B). In the sequence feature extraction, the raw Nanopore reads are first basecalled using Guppy (ONT Guppy, 2021), and then re-squiggled by Tombo (Stoiber *et al.*, 2016) to aligning the currents to the reference genome. Following Ni *et al.* (2019) median normalization (Stoiber *et al.*, 2016), was then applied to the raw signals. To obtain the error features, variants need to be called from fastq files by samtools (Li *et al.*, 2009) after the reference alignment using minimap2 (Li, 2018). The resulting files consist of the status of a position in a read (Match, Mismatch, Deletion, Insertion) and the quality of the call. This data is then processed to obtain the information of the presence or absence of mismatches, deletions and insertions at the genomic position of the read.

Specifically, features for DeepMP are selected through an Incremental Feature Selection (IFS) strategy. We applied this strategy to a set of 11 features. Seven of them for the sequence module: mean, median, standard deviation, range, skewness, kurtosis and the number of signals; and four for the error module: base quality, mismatches, deletions and insertions. The features are ranked by the model performance (Supplementary Figs S5 and S6) on one subset of the *E.coli* data and a subset of the human data (Supplementary Fig. S7). Both

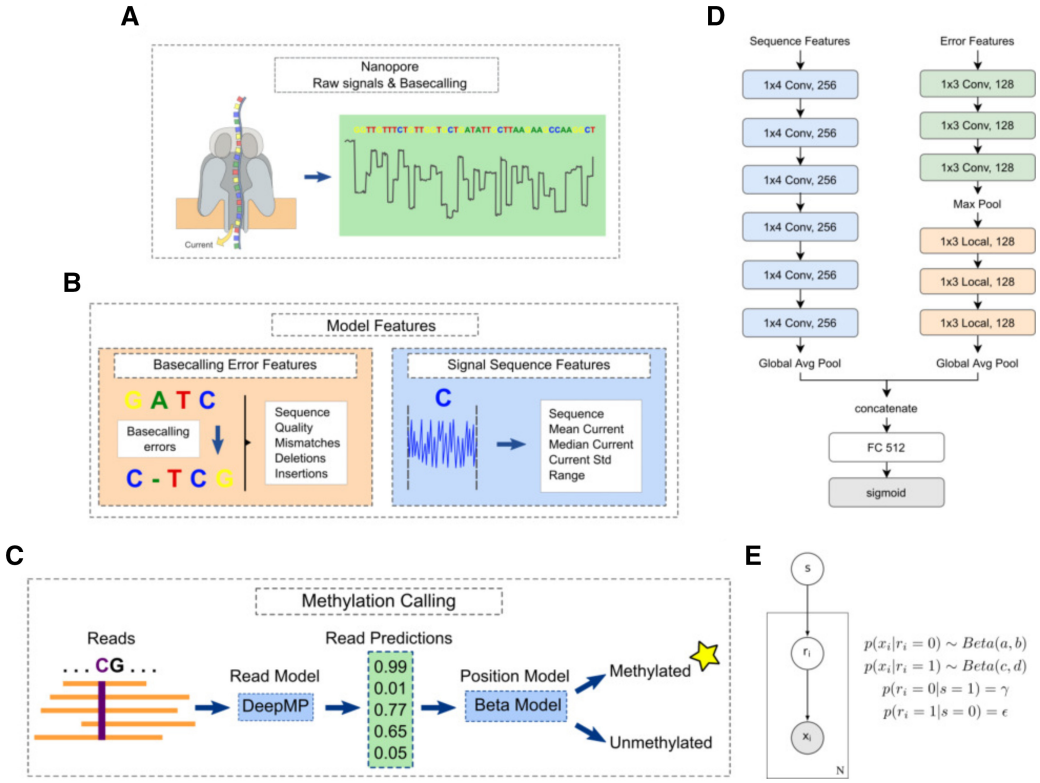

Fig. 1. DeepMP model overview. (**A**) Scheme of Nanopore technology, raw signals and basecalling. (**B**) Input features are divided into the basecalling error features (left) and the signal sequence features (right). Error features comprise the sequence information, read quality, mismatches, deletions and insertions. Sequence features consist of the sequence data and the mean, median, SD and range of the base currents. (**C**) Methylation calling pipeline. (**D**) DeepMP architecture. The sequence module involves 6 1D convolutional layers with 256 1×4 filters. The error module comprises 3 1D convolutional layers and 3 locally connected layers both with 128 1×3 filters. Outputs are concatenated and inputted into a fully connected layer with 512 units. (**E**) Probabilistic graphical model of the proposed Bayesian approach for position-based calling. The input data x_i consists of N read predictions. r_i is the read state, and s is the position state to be detected. Shaded nodes represent observed values, while the values of unshaded nodes are inferred

sample sets contain 100 000 labeled examples. The selection starts with the top 1 ranked feature (signal mean in *E.coli*) and iteratively adds features into the combination according to the rank until the end of the list is reached. If a decrease in performance is observed, the current feature will be excluded in the following run. A 5-fold cross-validation was used to evaluate the performance of the selected features (Supplementary Figs S5–S7).

Based on this strategy, the features used by DeepMP are the mean, median, standard deviation and value range of the Nanopore signals for the sequence module and the read quality, deletions, insertions and mismatches for the error module. After feature extraction for each target base, an l-length vector is constructed for every feature. Every vector contains the value of the nucleotide of interest and its l - 1 closest neighbors from both directions.

2.4 DeepMP Framework

Recent efforts to detect modifications (Liu *et al.*, 2019b; Ni *et al.*, 2019) for Nanopore sequencing data have been exploring long short-term memory recurrent neural networks (LSTM RNNs), which are neural network models designed to learn long-term dependencies in sequential data. On a high level, our prediction task could be regarded as a many-to-one problem where the input of the model is a sequence and the output is a single value indicating the methylation state of the targeted base, which appears to naturally match the specialty of RNNs. However, this perspective overlooks some characteristics of the data: first, the inputs of the mentioned models are statistical measures of the signals instead of the raw signals, which have weaker time dependencies between the variables. Second, the base in interest locates at the middle of the input sequence rather than the end. In the mentioned methods, a bidirectional RNN first processes the sequence from one end to the other to output a representation for the whole sequence, then performs the same process for the reversed sequence. The representations obtained from this procedure are most sensitive to the input bases located at both ends of the sequence (Goodfellow *et al.*, 2016) instead of the middle one. Besides, since these methods use a fixed-length sequence as the input to the networks, the flexibility of RNNs to adapt to various sequence lengths is not fully utilized in this particular problem setting. In addition, the large amount of parameters in RNNs makes training the networks costly.

From a different aspect, one might consider the excerpted set of features to be a series of buckets containing information for each base in the sequence. The goal is to capture the interactions between the center bucket and its neighbors, as well as the interactions among them. Concerning the spatial correlation in these buckets, we propose to use convolutional neural networks (CNNs). By virtue of sparse connections between units and the parameter sharing in the convolutional layers, CNNs are memory efficient and easy to parallelize, therefore, less expensive to train compared with RNNs. CNNs have been successfully applied to numerous tasks such as classification, detection, segmentation in various study fields including computer vision, natural language processing, drug discovery, etc. In our DeepMP model, there are two CNN modules (Fig. 1D): the sequence module and the basecalling error module. They are designed to process two different sets of features extracted from Nanopore sequencing data.

2.4.1 Sequence module

The four l-length vectors generated by sequence feature extraction are stacked with the one-hot embedded nucleotide sequence to form the model input vector of size $l \times 9$. The input vector is then given to the sequence CNN module, which is composed of 6 1D

convolutional layers with 256 1×4 filters. The stride for convolution is fixed to 1 base. The convolution function computes the nth element Z on the feature map by

$$Z_n = \sum_j X_{(n-1)+j} K_j \tag{1}$$

where X is the input to the layer, K_j is the jth element in kernel tensor.

Batch normalization (BN) is applied to each convolutional layer right before a ReLU activation function. The sequence module ends with a global average pooling layer.

2.4.2 Basecalling error module

With the same treatment as to the sequence feature, the l-length vectors are stacked with the one-hot embedded nucleotide sequence, resulting in the final $l \times 9$ $l \times 9$-shaped input for the error module.

The error module contains two types of layers: 1D convolutional layers and locally connected layers (LeCun, 1989). The convolutional layers compress the features into a compacted representation whereas the locally connected layers detect the neighborhood information in the sequence. A locally connected layer learns a set of filters separately for every location, naturally, it captures spatial characteristic information of the features. Since the signals from one base in the DNA sequence is highly correlated with its close neighboring bases, learning the spatial patterns enhances the model's ability to discriminate between different modification states. The input vector is fed into a three-layer convolutional neural network followed by a max-pooling layer, and then a three-layer locally connected network, finally a global average pooling layer. Both of the convolutional layers and locally connected layers have 128 1×3 filters.

2.4.3 Model outputs

The outputs from the sequence module and the error module are concatenated and inputted into a fully connected layer with 512 units. Later on, the last fully connected layer with a sigmoid activation function outputs the final prediction $\hat{y} \in [0, 1]$ for the central base of the read. This target position is called to be methylated if the model outputs a value ≥ 0.5.In addition, the two modules can be independent models by themselves.

2.4.4 Training DeepMP

During the training process, the model learns to minimize the loss computed by the binary cross-entropy loss function

$$\text{Loss} = -\frac{1}{m}\sum_{i=1}^{m} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i) \tag{2}$$

where y is the label, $\hat{y}$ is the model output, and m is the mini-batch size.

We trained the model using an Adam optimizer with a learning rate of 0.00125 and a mini-batch size of 512, early stopping was applied to prevent overfitting. In the default settings, we set the sequence length l to be 17. The full implementation is in Python 3 with Tensorflow 2.

In the experiments, DeepMP was trained for 10–20 epochs depending on the size of the dataset. On *E.coli* we performed a 5-fold cross-validation following the strategy proposed by Liu *et al.* (2019b). The *E.coli* genome was split into five sections ([0, 1000000], [1000000, 2000000], [2000000, 3000000], [3000000, 4000000], [4000000, 4700000]) (Supplementary Table S2). Then, models were trained on four sets and test on the remaining one. For the human dataset, chromosome 1 was kept for testing, while other regions were used for training and validation. For the pUC19 plasmid DNA data, reads were split into a 90% training set, a 5% validation set and a 5% test set. In all datasets, training, test and validation sets contained 50/50 positive and negative samples.

2.4.5 Position modification calling

Individual read-based calls mapped to a certain position in the genome need to be gathered to predict the methylation status of that position. Particularly, we introduce a Bayesian approach to accomplish the position modification calling (Fig. 1E). Once the neural networks are tested on labeled data, read-based predictions for each label group can be obtained. These values are subsequently used to infer the parameters of the underlying distribution for each group. By incorporating the read-based predictions and the ground truth labels, the proposed Bayesian approach provides more precise calls for the genomic positions.

The Bayesian model is based on the following: let $\mathbf{x} = \{x_1, x_2, \ldots, x_N\} \in \Omega^N$ to be the prediction of N reads for one position. Assuming different read calls are independent and identically distributed, the probability of observing $\mathbf{x}$ given the position state s can be computed by

$$p(\mathbf{x}|s) = \prod_i \sum_{r_i} p(x_i|r_i) p(r_i|s) \tag{3}$$

where r_i indicates if the ith read mapped to the position is modified, $r_i \in \{0, 1\}$. The graphical model is shown in Figure 1E.

According to Bayes' rule, given a prior distribution $p(s)$ the posterior probability of s is in proportion to $p(\mathbf{x}|s)$

$$p(s|\mathbf{x}) = \frac{p(\mathbf{x}|s)p(s)}{p(\mathbf{x})} \propto p(\mathbf{x}|s)p(s) \tag{4}$$

To infer the position state, we need to compute $p(\mathbf{x}|s = 0)$ along with $p(\mathbf{x}|s = 1)$ and compare these two likelihoods. Notice that given position state s, the random variable r_i follows a Bernoulli distribution. Therefore, we model $p(r_i|s)$ with two Bernoulli distributions parameterized by ϵ and γ:

$$\begin{cases} r_i \sim Bern(\epsilon), & \text{if } s = 0 \\ r_i \sim Bern(1 - \gamma), & \text{if } s = 1 \end{cases} \tag{5}$$

then model $p(x_i|r_i)$ with beta distribution parameterized by $a, b, c, d \sim \{z \in \mathbb{R} | z \geq 0\}$

$$\begin{cases} x_i \sim Beta(a, b), & \text{if } r_i = 0 \\ x_i \sim Beta(c, d), & \text{if } r_i = 1 \end{cases} \tag{6}$$

with Equations 3–6, we arrive at the final expression:

$$p(\mathbf{x}|s) = \prod_i \begin{cases} (1 - \epsilon) \cdot \mathbf{f}(x_i; a, b) + \epsilon \cdot \mathbf{f}(x_i; c, d), & \text{if } s = 0 \\ \gamma \cdot \mathbf{f}(x_i; a, b) + (1 - \gamma) \cdot \mathbf{f}(x_i; c, d), & \text{if } s = 1 \end{cases} \tag{7}$$

where $\mathbf{f}$ is the probability density function of beta distribution. Once $p(\mathbf{x}|s = 1)$ and $p(\mathbf{x}|s = 0)$ are computed, the methylation state of the position is decided by the higher probability among them. For example, a state is called positive, i.e. methylated when $p(\mathbf{x}|s = 1) \geq p(\mathbf{x}|s = 0)$ and negative otherwise.

We now discuss how to choose parameters γ, ϵ and a, b, c, d. By having γ close to 1.0 and ϵ close to 0, the model becomes prone to call positions methylated at low frequency as methylated. In the experiment, we fix $\gamma = 0.83$ and $\epsilon = 0.05$. The parameters for the beta distribution are estimated from the read-based predictions on an independent labeled dataset. The predictions are divided into two groups according to the label of the sample, where we estimate the sample mean μ and variance for each group. The shape parameters α and β for beta distribution can be approximated by

$$\begin{cases} \alpha = \left(\dfrac{1 - \mu}{var} - \dfrac{1}{\mu}\right)\mu^2 \\ \beta = \alpha\left(\dfrac{1}{\mu} - 1\right) \end{cases} \tag{8}$$

finally, we obtain two sets of α and β: $\{a, b\}$ for label 0 group and $\{c, d\}$ for label 1 group.

2.4.6 Performance evaluation

We quantify the performance of the models by four common measures: accuracy, precision, recall and F-score. The ground truth labels (positive and negative) are obtained from the bisulfite sequencing for the human data, as described in Section 2.1.2; for *E.coli* data, they are obtained as described in Section 2.1.1. We define an instance as true positive (TP) or true negative (TN) when the model prediction is consistent with the ground truth label in methylated and unmethylated examples, respectively. A false-positive (FP) instance corresponds to the model predicting a methylated site when the ground truth is labeled unmethylated. In contrast, a false-negative (FN) instance results when a model classifies a methylated site as unmethylated.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{F} - \text{Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{12}$$

$$\text{Methylated Frequency} = \frac{\sum \text{Methylated Reads}}{\text{Total number of reads}} \tag{13}$$

where TP, TN, FP, FN, are abbreviations for true positives, true negatives, false positives and false negatives, correspondingly. The accuracy gives an overview of the quality of the predictions. The precision shows the correct predictions in the positive calls, and recall is the percentage of correct predictions in all positive samples. In addition, the F-score is the harmonic mean of precision and recall, which shows the balance between the two. Finally, the methylation frequency of a position is calculated as the quotient between the number of methylated reads and the total number of reads overlapping the position.

3 Results

DeepMP takes as input two types of information from Nanopore sequencing data (Fig. 1A), basecalling errors and raw current signals (Fig. 1B). Features from these two types of information are fed into a CNN-based module to identify modified sites in individual reads (Fig. 1D). These read-level predictions are then integrated by the position-based calling method described in Section 2.4.5 (Fig. 1E), to identify methylated genomic sites (Fig. 1C).

We assessed the performance of DeepMP in the task of detecting methylated sites on three different datasets, *E.coli*, human and pUC19. The *E.coli* dataset has been widely used for model training and benchmarking. It consists of a negative control with 0% methylation (PCR amplified) and a positive one treated with a methyltransferase (M.SssI) that converts 5C to 5mC with a ~95% efficiency (Simpson *et al.*, 2017) (Section 2.1.1). Regarding the human dataset, the available bisulfite sequencing allowed the labeling of high-quality methylated and unmethylated positions (Section 2.1.2). Moreover, the pUC19 dataset opens up the possibility to evaluate DeepMP in the identification of a different modification (6 mA) within GATC motifs. These three datasets were used to benchmark DeepMP against Megalodon (ONT Megalodon, 2021), Guppy (ONT Guppy, 2021), DeepSignal (Ni *et al.*, 2019) and Nanopolish (Simpson *et al.*, 2017).

3.1 5mC detection performance on *E.coli* data

DeepMP and DeepSignal were trained on a mixture of methylated and unmethylated reads as described in Section 2.4.4. For the other methods, we used the available pre-trained models and the procedures explained in Section 2.2. All methods were tested on the same set of the data.

3.1.1 Prediction accuracy and studies at read level

To determine whether a genomic position is methylated, the first step is to detect the methylation status of each read covering it (Fig. 1C). Thus, an independent set of reads, as described in Section 2.4.4 is selected to evaluate all three models. DeepMP showed the best Area under the ROC (AUC) among all the benchmarked methods (DeepMP: 0.988, Megalodon: 0.981, DeepSignal: 0.974, Guppy: 0.924, Nanopolish: 0.878) (Fig. 2A). While Guppy and Megalodon presented the highest precision values (Guppy: 0.9957, Megalodon: 0.9906), DeepMP exhibited the highest overall accuracy (0.9397), recall (0.9347) and f-score (0.9398) (Fig. 2B). These results were consistent throughout the 5-fold cross-validation (Fig. 2C; Supplementary Table S2).

These read predictions come from a mixture of two samples that are completely methylated or unmethylated. However, in a natural population of cells, one particular base may be methylated across some, but not all of them. For instance, 6mA across 12 different genomic sites is found at levels ranging from 7% to 69% in yeast samples (Garcia-Campos *et al.*, 2019). To benchmark DeepMP in a real-life scenario, we thus generated 11 synthetic datasets featuring mixtures of methylated and unmethylated reads at different proportions. The methylated frequency of a given position is estimated by the ratio of predicted methylations to the total number of reads mapping that position. The accuracy was measured as $1 - L1_{\text{meth frequency}}$. The $L1_{\text{meth frequency}}$ distance is defined as the absolute value of the difference between the true methylated frequency of the position and that estimated by the model. This measurement takes into account FP and FN at read level, avoiding biased estimates of the methylated frequency. Figure 2D shows the improved performance of DeepMP for inferring the true methylated frequency of the sample. The accuracy of DeepMP was the highest among the methods at high proportion of methylated reads. At low proportions of methylated reads, Megalodon and Guppy showed a comparable performance. These results are in accordance with the FN and FP levels shown in Figure 2E and Supplementary Figure S1A. DeepMP presents a lower number of FN across all proportion of methylated reads, and the levels of FP are consistent across synthetic datasets for all methods except Nanopolish.

One of the main innovations of DeepMP consists in combining a basecalling error module and a sequence signal module. To assess how much this innovation actually contributes to the observed gain in performance, DeepMP trained on the sequence module only (DeepMP Seq) was also included in the comparison (Supplementary Fig. S1C and D). DeepMP Seq shows a decreased performance compared with DeepMP (DeepMP F-score: 0.9398; DeepMP Seq F-score: 0.889) (Supplementary Table S2), which highlights the importance of taking into account the features derived from basecalling errors for the *E.coli* dataset.

3.1.2 Prediction accuracy on genomic sites

The ultimate goal of methods that identify DNA methylation is to be correctly determine the frequency of methylation of a genomic site across cells. This is done through the identified read level methylation sites of reads overlapping the genomic position under analysis. One strategy to do so is to decide a particular percentage of methylated reads as the threshold (Liu *et al.*, 2019b), i.e. a 20% threshold will detect as methylated any position presenting more than 20% of methylated overlapping reads. Another approach is to get a mean estimate of the predictions of the n reads overlapping a particular position (Ni *et al.*, 2019).

Although such approaches yield acceptable results on datasets with sites methylated at high-frequency (Liu *et al.*, 2019b; Ni *et al.*, 2019), their accuracy usually drops at sites methylated at lower levels (10–30%) (Fig. 2F) (a 20% threshold was applied to all models but DeepSignal and DeepMP). To solve this problem, we propose to apply a Bayesian approach (DeepMP; Figs 1E and 2F), which utilizes the statistical features of the read predictions to compute the likelihood of the different states for the genomic position (Section

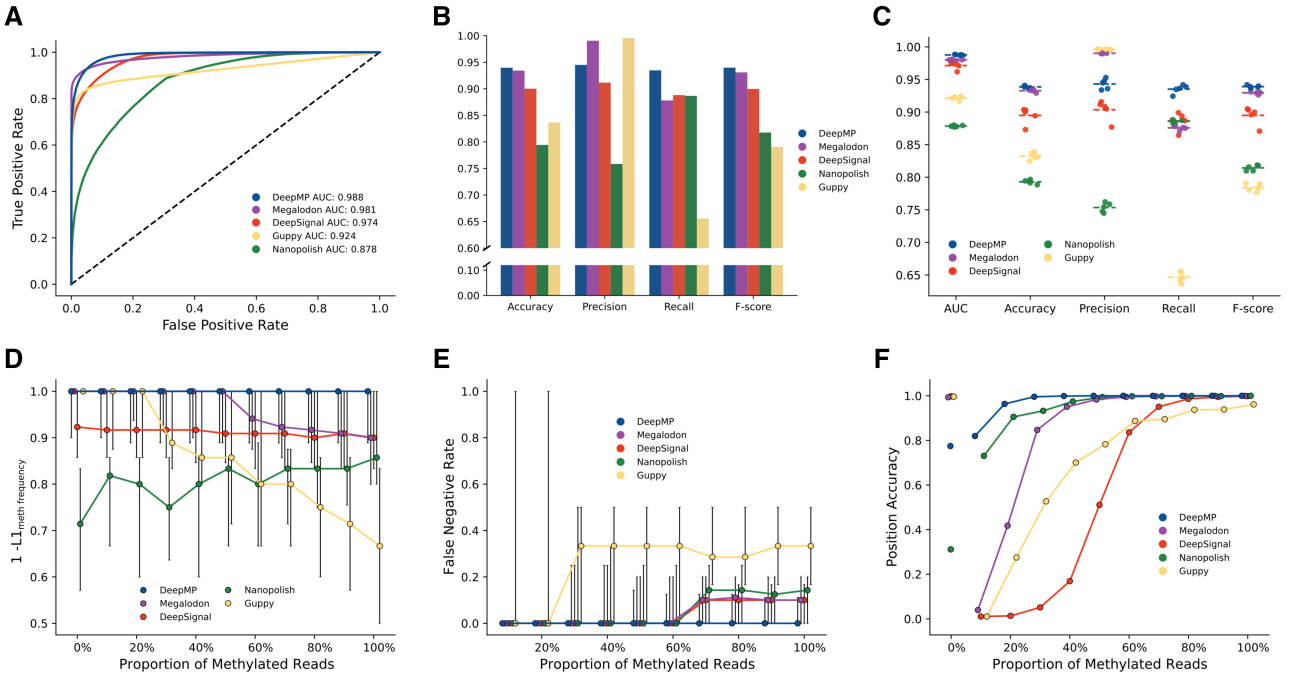

Fig. 2. Performance of DeepMP, Megalodon, Guppy, DeepSignal and Nanopolish on *E.coli* dataset. (**A**) Receiver operating characteristic (ROC) curve showing the false-positive rate (*x*-axis) versus the true-positive rate (*y*-axis) for the read predictions on a mixture of methylated and unmethylated reads in a single cross-validation fold. (**B**) Accuracy measurements for the compared models in the cross-validation fold in A. (**C**) 5-fold cross validation accuracy and AUC results for mixtures of methylated and unmethylated reads. (**D**) Accuracy to determine the methylation frequency of a sample measured by 1 - $L1_{\text{meth frequency}}$ (*y*-axis). 1 - $L1_{\text{meth frequency}}$ is evaluated on 11 datasets comprising different levels of methylated reads (0–100%) (*x*-axis). (**E**) False-negative rate (*y*-axis) evaluated on the same 11 datasets as in D. Dots in E and D represent the median observation and black lines the first (Q1) and third (Q3) quartiles. (**F**) Position calling accuracy for each of the partially methylated datasets at different thresholds (DeepMP: Bayesian approach; DeepSignal: mean read predictions estimate; Megalodon, Guppy and Nanopolish: 20% threshold). Note that, jitter has been added to D, E and F to avoid overlapping of dots in the graph

2.4.5). As a result, DeepMP shows the highest accuracy (>80% of the positions) when the dataset contains, on average, sites methylated in 10% of the reads (Fig. 2F), while robustly calling ~80% of the positions when the dataset is completely unmethylated. This trend is conserved when comparing DeepMP using different thresholds (10%, 20% and 50% Threshold) (Supplementary Fig. S1B).

3.2 5mc detection performance on human data

DeepMP and DeepSignal were trained on reads containing methylated and unmethylated cytosines (determined by bisulfite sequencing) and tested on the same type of data drawn from reads overlapping human chromosome 1 in the Norwich subset of the human data (Supplementary Table S1). For Megalodon, Guppy and Nanopolish, we used available pre-trained models as explained in Section 2.2 and we tested them on the same set of data.

3.2.1 Prediction accuracy at read level

Similar to *E.coli* dataset, DeepMP, Megalodon, Guppy, DeepSignal and Nanopolish were evaluated on their read prediction performance at read level (Fig. 3A and B; Table 1). DeepMP achieves the better ROC AUC scores (DeepMP: 0.967; Megalodon: 0.9394; Guppy: 0.8602; DeepSignal: 0.9629; Nanopolish: 0.9284) and F-scores (DeepMP: 0.9324; Megalodon: 0.8968; Guppy: 0.7787; DeepSignal: 0.9255; Nanopolish: 0.9236).

When comparing DeepMP Seq (DeepMP trained only using the sequence module) against DeepMP, Megalodon, DeepSignal, Guppy and Nanopolish, it outperforms all methods but DeepMP (DeepMP Seq AUC: 0.966; DeepMP Seq F-score: 0.9315) (Supplementary Fig. S2A and B; Supplementary Table S3). These results differ from the *E.coli* dataset where DeepMP Seq is behind DeepSignal in performance measurements.

3.2.2 Correlation with bisulfite sequencing and biologically meaningful regions

To further evaluate the ability of models to infer the correct frequency of methylation of methylated positions, we evaluated the correlation of their estimated frequency with the result obtained from the bisulfite sequencing data. We computed this correlation across all probed genomic sites. Moreover, we estimated the performance of all methods to call methylated sites overlapping two types of functionally relevant genomic regions (LINE1 genomic regions and genomic imprinting genes). DeepMP (shown in Fig. 3C) showed the highest Pearson's correlation value ($r = 0.866$) and coefficient of determination ($r^2 = 0.75$), as well as the lowest Root Mean Square Error (RMSE = 0.204) compared with DeepSignal ($r = 0.843$, $r^2 = 0.711$, RMSE = 0.218), Megalodon ($r = 0.809$, $r^2 = 0.655$, RMSE = 0.202), Nanopolish ($r = 0.850$, $r^2 = 0.723$, RMSE = 0.214) and Guppy ($r = 0.599$, $r^2 = 0.359$, RMSE = 0.395) at the genome-wide scale.

We then assessed the performance of the methods to detecting methylation at sites overlapping LINE1 regions (which tend to be repetitive) or imprinting genes. At read level, positions were selected based on the criteria explained in Section 2.1.2 to generate ground truth labels. Supplementary Figure S4A and B shows that DeepMP outperforms other methods in terms of AUC (0.968), accuracy (0.917) and F-score (0.926) in the identification of methylation sites at imprinting genes. DeepMP also outperforms other methods (AUC: 0.945; accuracy: 0.896; F-score: 0.896) in the identification of methylated sites overlapping LINE1 regions (Supplementary Fig. S4C and D).

Finally, we also compared the correlation of the methylation frequency estimated by the methods in the two latter datasets (LINE1 and imprinting genes) with the results from bisulfite sequencing (shown in Fig. 3F alongside the aforementioned values for genome-wide sites). DeepMP shows the highest correlation in the three datasets (LINE1: 0.737; Imprinting genes: 0.895; Genome-wide: 0.866).

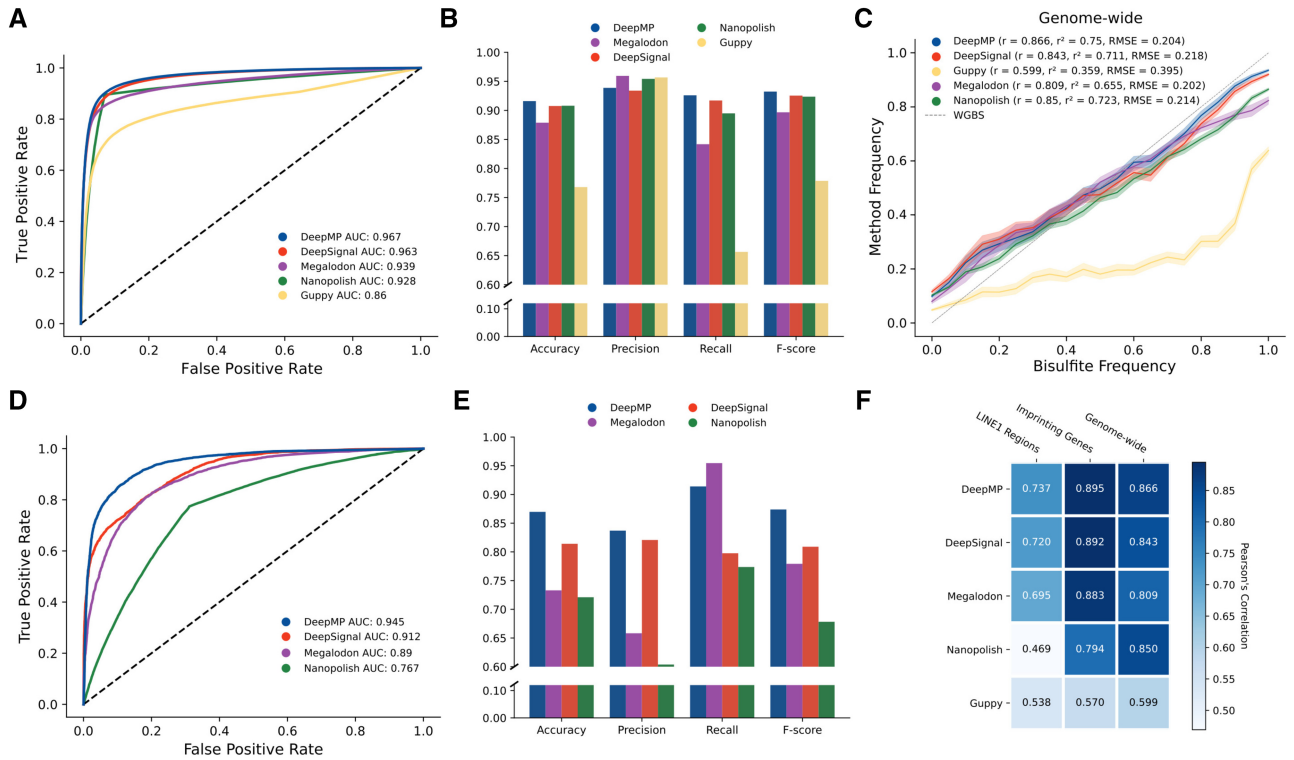

Fig. 3. Performance of DeepMP, Megalodon, Guppy, DeepSignal and Nanopolish on the Norwich subset of the human dataset and of DeepMP, Megaldon, DeepSignal and Nanopolish on the pUC19 plasmid. (**A, D**) Receiver operating characteristic (ROC) curve showing the false-positive rate (*x*-axis) versus the true-positive rate (*y*-axis) for the read predictions on a mixture of methylated and unmethylated reads on the human dataset (A) and pUC19 (D) respectively. (**B**) Accuracy measurements for the models on the human dataset. (**E**) Accuracy measurements for the models on the pUC19 plasmid. (**C**) Correlation between the methylation frequency of a method and bisulfite sequencing for the human dataset. Lines denote the mean methylated frequency of a method for a given bisulfite frequency and the shaded surface corresponds to the 95% confidence interval. (**F**) Pearson's correlation of every method for three different genomic regions (LINE1, Imprinting genes, Genome-wide) of the human dataset

Table 1. Performance summary at read level of DeepMP, Nanopolish, DeepSignal, Guppy and Megalodon on *E.coli*, human and pUC19 datasets

Dataset	Test	Method	Accuracy	Precision	Recall	*F*-score
E.coli	Read level	Nanopolish	0.7929	0.7535	0.8859	0.8143
		DeepSignal	0.8948	0.9035	0.8865	0.8949
		Guppy	0.8322	**0.9961**	0.6466	0.7841
		Megalodon	0.9327	0.9901	0.8757	0.9294
		DeepMP	**0.9386**	0.9431	**0.9352**	**0.9391**
Homo sapiens	Read level	Nanopolish	0.9080	0.9543	0.8949	0.9236
		DeepSignal	0.9076	0.9340	0.9191	0.9255
		Guppy	0.7680	0.9568	0.6565	0.7787
		Megalodon	0.8788	**0.9594**	0.8418	0.8968
		DeepMP	**0.9160**	0.9388	**0.9260**	**0.9324**
pUC19 plasmid	Read level	Nanopolish	0.7211	0.6038	0.7737	0.6783
		DeepSignal	0.8141	0.8207	0.7976	0.8090
		Megalodon	0.7730	0.6583	**0.9547**	0.7792
		DeepMP	**0.8697**	**0.8370**	0.9141	**0.8738**

Note: Accuracy measurements for *E.coli* represent the average value of the 5-fold cross-validation. Bold entries represent the largest measurement value for every data set.

Furthermore, to study the correlation with bisulfite more in detail, positions were divided into three groups based on the level of methylation in bisulfite sequencing (low, intermediate and high). Megalodon and Guppy show higher performance at sites methylated at low frequencies. In contrast, DeepMP and DeepSignal exhibit better performance (and consistent results across the three categories) with bisulfite sequencing at intermediately and highly methylated sites in imprinting genes (Supplementary Fig. S3A) and LINE1 (Supplementary Fig. S3B).

3.3 6mA detection performance on pUC19 plasmid DNA data

To evaluate the ability of DeepMP to detect a different type of methylation (6mA), we benchmarked DeepMP on the pUC19 dataset. The read-based strategy was used to train DeepMP and DeepSignal on a mixture of methylated (6mA) and unmethylated (A) reads, as described in Section 2.4.4. As the model selected for Guppy does not specifically detect modifications in all contexts (Section 2.2), it was discarded. Nevertheless, Megalodon, considered (and supported by our previous results) the state-of-the-art Nanopore method for methylation detection (ONT Megalodon, 2021; Yuen *et al.*, 2021), rather than Guppy, was run with the available model for all contexts as described in Section 2.2. Nanopolish was run to detect modifications in *dam* sites (DNA adenine methylase). Considering the number of reads discarded by the selective cutoff, we tested Nanopolish on all the extracted examples. The rest of the models used the same set of data for testing.

DeepMP significantly outperforms DeepSignal, Megalodon and Nanopolish in the detection of 6mA within GATC motifs (Table 1). In Figure 3D and E, DeepMP also shows an improvement of performance in the binary classification task and the related accuracy measurements. Furthermore, DeepMP Seq alone was also able to outperform any other method but DeepMP in the detection of modified GATC motifs (Supplementary Fig. S2C and D; Supplementary Table S3).

4 Discussion

Nanopore sequencing, coupled with the computational methods that exploit its output to detect base modifications, has opened up the possibility of directly identifying epigenetic modifications of DNA nucleotides. The advent of this technology has the potential to supersede methodologies like bisulfite sequencing, which is currently the gold standard to detect DNA methylation. For this to be accomplished, the accuracy computational methods of detection need to be improved to correctly identify the methylation level at individual genomic positions (Garcia-Campos *et al.*, 2019). This study proposes a neural network solution, DeepMP, which utilizes both electric currents and basecalling errors from Nanopore sequencing to deliver high accurate methylation detection.

We showed that DeepMP outperforms state-of-the-art methods Megalodon, Guppy, DeepSignal and Nanopolish on *E.coli*, human and pUC19 datasets. DeepMP favorably compared in the proposed benchmark when inferring the true methylated frequency of partially methylated *E.coli* datasets (0%, 10%, …, 100%), consistently presenting lower false positive and negative rates. DeepMP also outperformed other methods in the estimated frequency of methylation across sites in the human data. Furthermore, the analysis of the methods' ability to call the position methylation status displayed the limitation of introducing an arbitrary, human-defined threshold. A 20% threshold could not properly capture a position's true methylation status in datasets with low methylation levels (<30% methylation). Importantly, using DeepMP's novel Bayesian approach corrected this problem while being robust on completely methylated samples.

Intriguingly, in the case of the human dataset, the features based on basecalling errors do not appear to provide the same level of information as in the *E.coli* dataset. This finding indicates that the informational value of basecalling errors varies in different types of data, i.e. the error features may be dataset-specific. A hypothesis would be that the M.SssI treatment in *E.coli* samples generates a characteristic error pattern that is not present when mutations occur naturally (human dataset). Consequently, to obtain a more generalizable model across species, one might consider using DeepMP without the basecalling errors module.

Moreover, we demonstrated that DeepMP detects not only 5mC but also 6 mA more accurately than state-of-the-art methods. Given the low number of reads, DeepMP achieved a significant separation compared with Megalodon, DeepSignal and Nanopolish. This fact highlights the versatility of DeepMP across different types of nucleotide modifications (including at certain sequence motifs) and sequencing coverage. Nevertheless, it shares one of the current limitations of most supervised learning models, namely, the inability to detect out-of-sample modifications. That is, to detect methylated bases absent in the training set. One way to overcome could be to use unsupervised or one-shot learning methods (Koch *et al.*, 2015; Vinyals *et al.*, 2016). This technical advancement would allow the exploration of a different set of interesting problems regarding the identification of modified bases using Nanopore.

In summary, DeepMP accurately detects methylated sites both, in individual reads and at genomic positions, at different levels of methylation of the sample, as real-world biological samples. Besides, the proposed Bayesian approach could apply to related problems to substitute a human-defined threshold, especially when labeled data is available. In addition, we posit that DeepMP could also be used to generate genome-wide maps of different DNA base lesions from Nanopore sequencing data. This could largely reduce experimental hurdles to achieve this objective. As Nanopore technology continues to advance and the sequencing costs are reduced, the amount of data will exponentially increase. Accurate and efficient methods as DeepMP will be key to exploit this sequencing data in scenarios with limited computational resources.

Acknowledgements

IRB Barcelona is a recipient of a Severo Ochoa Centre of Excellence Award from the Spanish Ministry of Economy and Competitiveness (MINECO; Government of Spain) and was supported by CERCA (Generalitat de Catalunya). The authors thank Zaka Yuen for her generous assistance in basecalling *E.coli* reads with Guppy.

Funding

This work was funded by ITN-CONTRA EU [H2020 MSCA-ITN-2017-766030 to J.B. and M.C.].

Conflict of Interest: none declared.

References

Bergman,Y. and Cedar,H. (2013) DNA methylation dynamics in health and disease. *Nat. Struct. Mol. Biology*, **20**, 274–281.

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**, 461–465.

Garcia-Campos,M.A. *et al.* (2019) Deciphering the "m6a code" via antibody-independent quantitative profiling. *Cell*, **178**, 731–747.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep learning*. MIT press.

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

Grønbaek,K. *et al.* (2007) Epigenetic changes in cancer. *APMIS*, **115**, 1039–1059.

Jain,M. *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, **36**, 338–345.

Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, **13**, 484–492.

Koch,G., *et al.* (2015) Siamese neural networks for one-shot image recognition. *ICML Deep Learn. Workshop*, **2**.

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.

LeCun, Y. (1989) Generalization and network design strategies. *Connectionism in perspective*, **19**, 143–155.

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

Li,H. *et al.*; 1000 Genome Project Data Processing Subgroup. (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.

Liu,H. *et al.* (2019a) Accurate detection of m 6 a RNA modifications in native RNA sequences. *Nat. Commun.*, **10**, 1–9.

Liu,Q. *et al.* (2019b) Detection of DNA base modifications by deep recurrent neural network on oxford nanopore sequencing data. *Nat. Commun.*, **10**, 1–11.

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.* (2019) Single-molecule sequencing detection of n 6-methyladenine in microbial reference materials. *Nat. Commun.*, **10**, 1–11.

Miura,F. *et al.* (2012) Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res.*, **40**, e136.

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

ONT Guppy. (2021) Github.com/nanoporetech. Retrieved June 2021.

ONT Megalodon. (2021) *github.com/nanoporetech/megalodon*. Retrieved June 2021.

Rand,A.C. *et al.* (2017) Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods*, **14**, 411–413.

Schatz,M.C. (2017) Nanopore sequencing meets epigenetics. *Nat. Methods*, **14**, 347–348.

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

Schübeler,D. (2015) Function and information content of DNA methylation. *Nature*, **517**, 321–326.

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

Stoiber,M. *et al.* (2016) *De novo* identification of DNA modifications enabled by genome-guided nanopore signal processing. *BioRxiv*, 094672.

The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57.

Vinyals,O. *et al.* (2016) Matching networks for one shot learning. Advances in neural information processing systems, **29**, 3630–3638.

Wescoe,Z.L. *et al.* (2014) Nanopores discriminate among five c5-cytosine variants in DNA. *J. Am. Chem. Soc.*, **136**, 16582–16587.

Yuen,Z.W.-S. *et al.* (2021) Systematic benchmarking of tools for CPG methylation detection from nanopore sequencing. *Nat. Commun.*, **12**, 1–12.

Zhu,S. *et al.* (2018) Mapping and characterizing n6-methyladenine in eukaryotic genomes using single-molecule real-time sequencing. *Genome Res.*, **28**, 1067–1078.