Systematic benchmarking of tools for CpG methylation detection from Nanopore sequencing

Zaka Wing-Sze Yuen[1,2], Akanksha Srivastava[1,2], Dennis McNevin[3], Cameron Jack[2*], Eduardo Eyras[1,2,4,5*]

[1]EMBL Australia Partner Laboratory Network at the Australian National University, Acton ACT 2601, Canberra, Australia

[2]The John Curtin School of Medical Research, Australian National University, Acton ACT 2601, Canberra, Australia

[3]Centre for Forensic Science, School of Mathematical & Physical Sciences, Faculty of Science, University of Technology Sydney, Australia

[4]Catalan Institution for Research and Advanced Studies (ICREA), E08010 Barcelona, Spain

[5]Hospital del Mar Medical Research Institute (IMIM), E08003 Barcelona, Spain

* co-corresponding authors: cameron.jack@anu.edu.au, eduardo.eyras@anu.edu.au

Abstract

DNA methylation plays a fundamental role in the control of gene expression and genome integrity. Although there are multiple tools that enable its detection from Nanopore sequencing, their accuracy remains largely unknown. Here, we present a systematic benchmarking of tools for the detection of CpG methylation from Nanopore sequencing using individual reads, control mixtures of methylated and unmethylated reads, and bisulfite sequencing. We found that tools showed a tradeoff between false positives and false negatives, and presented a high dispersion with respect to the expected methylation frequency values. We described various strategies to improve the accuracy of these tools and proposed a new method, METEORE (https://github.com/comprna/METEORE), based on the combination of the predictions from two or more tools that has improved accuracy over individual tools. Snakemake pipelines are provided for reproducibility and to enable the systematic application of our analyses to other datasets.

DNA modifications play a fundamental role in genome stability and gene regulation during mammalian development, disease progression and aging[1-4]. Of more than 17 possible DNA modifications, the methylation of

cytosines at CG di-nucleotides (CpG), involving the addition of a methyl group ($-CH_3$) to the 5th carbon of the cytosine ring to form 5-methylcytosine (5mC) is the most frequently observed in relation to gene regulation[5]. Key advances in the understanding of the function of 5mC have been made possible through the development of dedicated genome-wide profiling techniques[4]. Commonly used methods include restriction enzyme digestion, affinity enrichment, or bisulfite conversion, followed by microarray hybridization or short-read sequencing[4]. While short-read sequencing has been very effective at mapping 5mC sites at genome-scale, it still presents various disadvantages including high mapping uncertainty in repetitive regions and amplification bias. Moreover, short-read approaches result in the loss of native biochemical modifications and involve the necessary coupling with lengthy protocols, such as a bisulfite conversion, which is known to degrade DNA[6]. The conversion of 5mC to uracil is also very sensitive to the reaction conditions[7].

In contrast, Nanopore long-read technologies provide many distinct advantages. They can sequence individual DNA molecules in their native state, harboring base modifications, without any prior enzymatic or chemical treatment, and without the need for PCR amplification[8, 9]. As a single DNA molecule travels through a nanopore, base modifications can be revealed by their unique signal shapes, which differ from the equivalent unmodified base[8-10]. However, signals are dependent on sequence context and different copies of the same molecule present considerable signal variation. Thus, it is necessary to apply computational models to interpret the signals and to predict the methylation status at a given CpG site. Multiple tools have been developed in recent years, but a lack of a systematic benchmarking poses a significant challenge for users in assessing the reliability of their predictions. Although Nanopore provides an unprecedented opportunity to detect methylation at the single-molecule level, its accuracy in this context remains largely unknown. This precludes the development of reliable and cost-effective applications in patient, forensic, and environmental samples. It is thus necessary to thoroughly characterize the strengths and limitations of the different available tools and establish optimal strategies for the accurate detection of 5mC in DNA.

We have performed a systematic benchmarking of six tools for 5mC detection from Nanopore sequencing fast5 files using individual reads, controlled methylation mixtures, Cas9-targeted sequencing, and whole genome bisulfite sequencing. We established their detection capabilities at the single-molecule level and in discerning different stoichiometries at different levels of coverage and GC-content. In general, the tested tools present a trade-off between true positives and false positives, and a high dispersion in the prediction of methylation frequencies. We propose various strategies to improve detection accuracy, including a novel consensus method, METEORE (https://github.com/comprna/METEORE), that combines the outputs from two or more tools.

Results

Detection of DNA cytosine methylation from Nanopore sequencing

We developed a standardized workflow to obtain 5mC calls at CpG sites from Nanopolish[8], Megalodon[11], DeepSignal[12], Guppy[13], Tombo[14] and DeepMod[15] (Fig. 1), which we have implemented in Snakemake pipelines[16] (https://github.com/comprna/METEORE). This workflow ensures consistent inputs and outputs for all tools and facilitates the integration and interpretation of DNA methylation. Nanopolish detects CpG methylation with a hidden Markov model (HMM), while both DeepSignal and DeepMod use neural networks, and Tombo applies a statistical test to identify DNA modifications. Both Tombo and DeepSignal re-squiggle the raw signals before detection. Differently from the other tools, Guppy basecalls 5mC directly using an extended alphabet. We additionally considered Megalodon, which identifies 5mC by anchoring the Guppy basecalling output to the reference and assigning a score for the candidate modified base. All tools output per-site methylation calls on both strands at genome level.

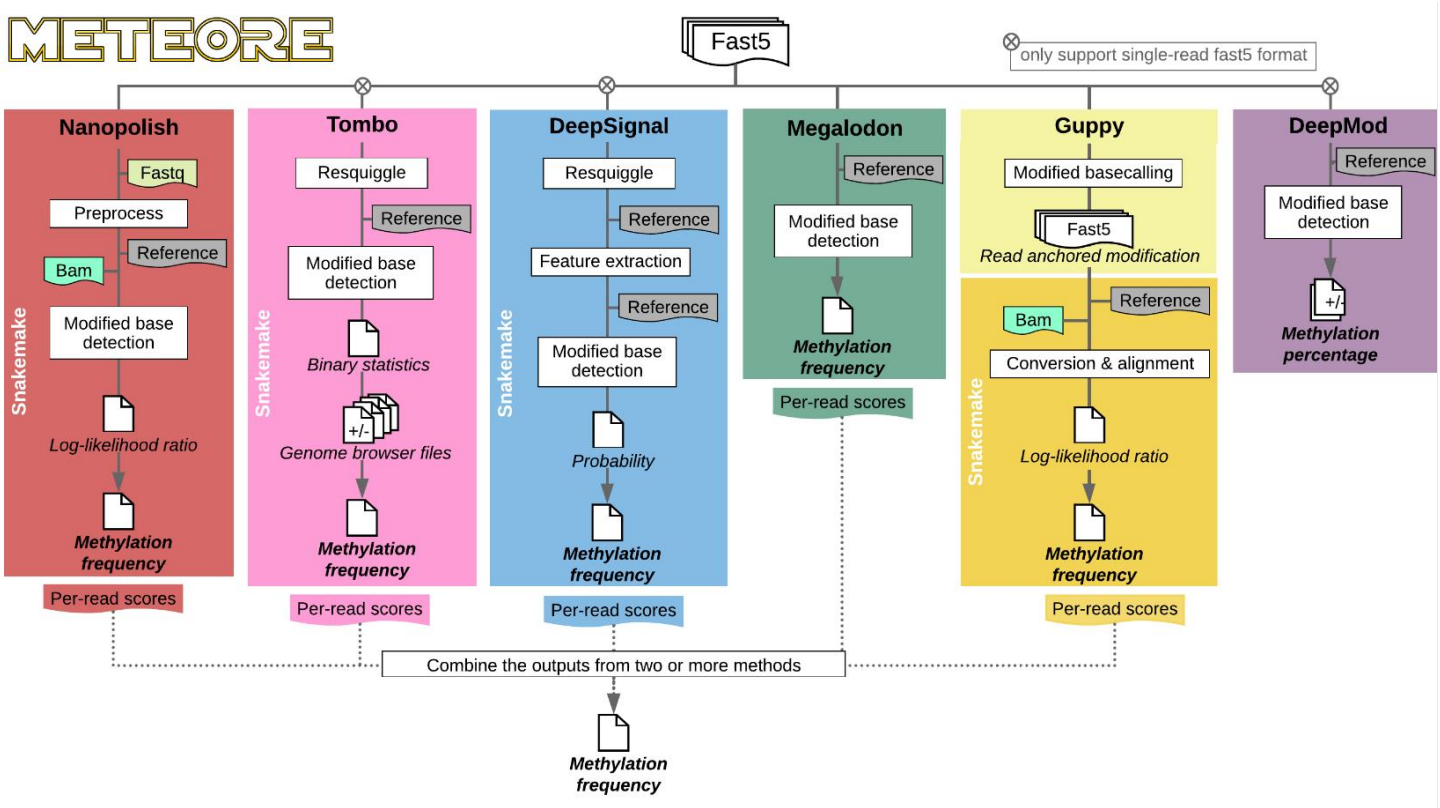

Figure 1. Analysis pipeline for 5mC detection at CpG sites from Nanopore sequencing. We describe the approach used to test the tools Nanopolish, Tombo, DeepSignal, Megalodon, Guppy and DeepMod. SnakeMake pipelines and command lines used are available at METEORE We did not develop a SnakeMake pipeline for Megalodon and DeepMod, as they were run with a single command. All tools, except for DeepMod, produce predictions per individual read and per CG site.

All tools make a prediction at genome level, which output a methylation frequency for each site, derived from fast5 input files. For Tombo, both methylation scores and coverage data were produced separately in WIG format (Genome browser files), which were then combined and converted into a format similar to the other tools. In addition to methylation frequency, tools except for DeepMod output a score to make a methylation call for each site at read level. We also indicate which methods currently only accept single-read fast5 format.

Performance of CpG methylation detection on methylation controls

We first evaluated the six tools on methylation control datasets built from PCR-amplified DNA (negative control) and M.SssI-treated DNA (positive control)[8]. From the 346,793 CpG sites in the *E. coli* reference genome, we selected 100 arbitrary sites containing a single CpG site in a 10nt window with minimum read coverage of 50x (median coverage 85x) from both positive (methylated) and negative (unmethylated) control datasets. Using these reads, we created 11 benchmarking datasets with specific mixtures of methylated and unmethylated reads, namely, containing 0%, 10%, …, 90%, and 100% of methylated reads, each set with approximately 2,400 reads (mixture dataset 1) (Supp. Table 1) (Supp. File 1). We examined the per-site methylation frequency predicted by each tool across these 100 sites using the default cutoff for each method. All tools, except Tombo, achieved Pearson correlations above 0.8 (p-value < 2.2e-16 for all tools) (Fig. 2a). The highest correlation was attained by DeepMod, closely followed by DeepSignal and Megalodon.

Despite the good correlation values, most tools showed high dispersion and low agreement with the expected percentage methylation per site. Guppy and Megalodon had the highest root mean square error (RMSE) values and systematically underpredicted per-site methylation (Fig. 2a). Nanopolish and Tombo showed high dispersion and consistently overpredicted. To further assess how the dispersion affected the per-site accuracy, we calculated the proportion of sites predicted outside a 10% window around the expected value for each percentage methylation subset. Megalodon and Guppy had most sites predicted outside expected windows (Fig. 2b). Nanopolish and Tombo had the lowest proportion of sites predicted outside the m90 and m100 windows but showed the highest proportions at low methylation frequency. In contrast, DeepMod and DeepSignal achieved the lowest RMSE values and had most sites predicted within the expected windows for all subsets (Fig. 2b).

We further assessed the classification of fully unmethylated and fully methylated sites according to specific thresholds (Supp. Table 2). For 0% methylation, Guppy and Megalodon correctly predicted zero methylation for most of the 100 sites (83 and 71 respectively) (Fig. 2c), and all were called correctly if accepting sites with proportions of 0.1 or less methylation. In contrast, all other methods only correctly predicted about half the sites (Fig. 2c). In the 100% methylated set, Nanopolish and Tombo correctly predicted most of the 100 sites as fully methylated using a cutoff of at least 0.8 (Fig. 2d). In contrast, Megalodon and Guppy failed to predict most of the

fully methylated sites at this cutoff (Fig. 2d). These differences in the accuracy at fully methylated and fully unmethylated sites, as well as a general high dispersion at intermediate methylation, motivated us to identify new strategies to achieve better accuracies.

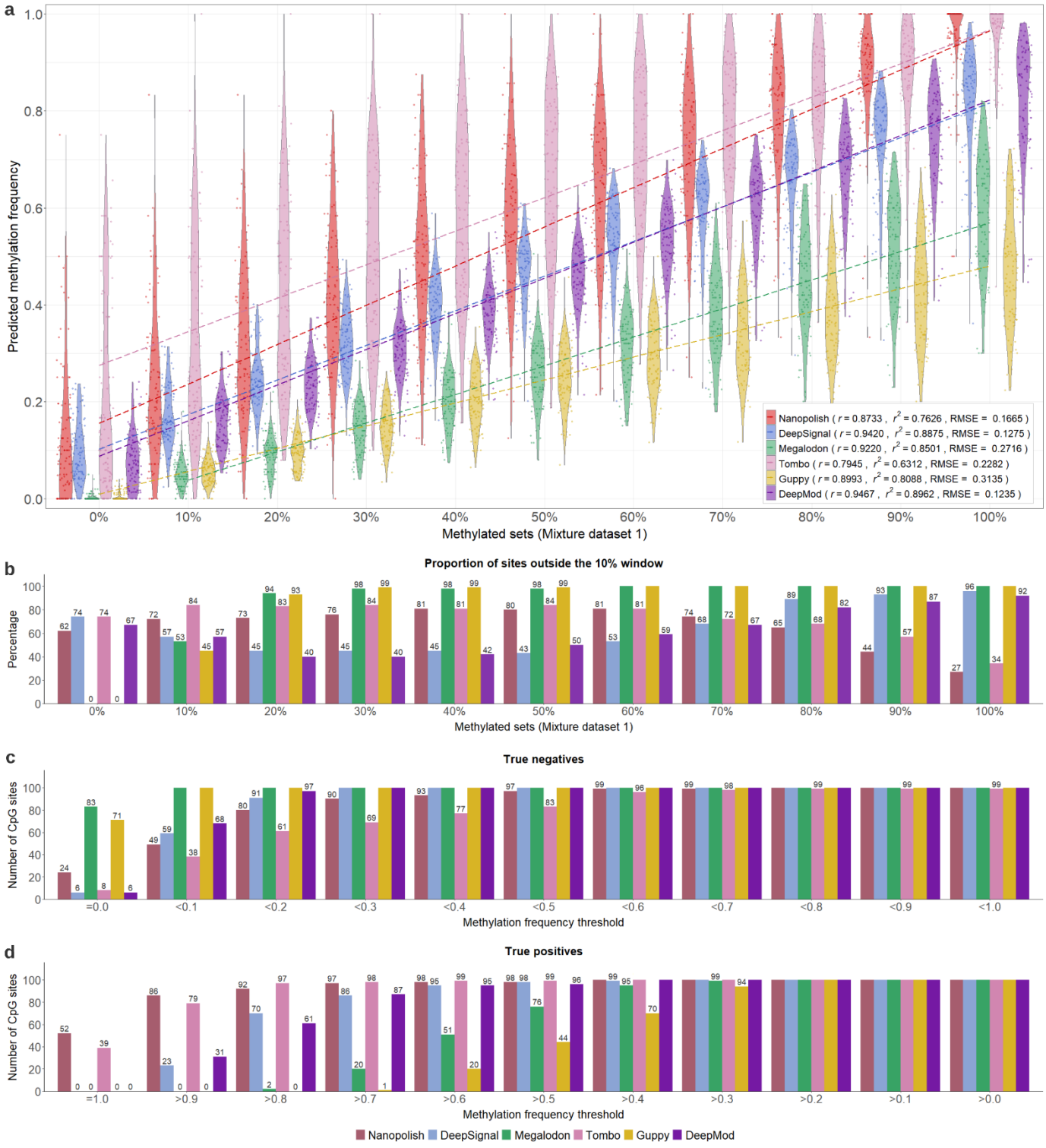

Figure 2. Accuracy analysis per CpG site on control mixture dataset 1. (a) Violin plots showing the predicted methylation frequencies (y axis) for each control mixture set with a given proportion of methylated reads (x axis). The

Pearson's correlation (r), coefficient of determination (r^2) and root mean square error (RMSE) are given for each tool. **(b)** For each method we indicate the proportion of sites predicted outside a 10% window around the expected methylation proportion, i.e. each predicted site in the m% dataset was classified as "outside" if its predicted percentage methylation was outside the interval [(m-5)%, (m+5)%] for intermediate methylation values, or outside the intervals [0,5%] or [95%,100%] for the fully unmethylated or fully methylated sets, respectively. We indicate the actual number on top of each bar, except when they reach 100%. **(c)** The number of true negatives (y axis) for each tool according to different thresholds for the predicted methylation frequency below which a site was called unmethylated (x axis), using the dataset of 100 fully unmethylated sites. Tombo did not produce any output for one of the tested sites. (**d)** The number of true positives (y axis) for each tool according to different thresholds for the predicted methylation frequency above which a site was called fully methylated (x axis), using the dataset of 100 fully methylated sites.

Methylation prediction accuracy in individual molecules

Nanopore sequencing provides the opportunity to detect nucleotides and their modifications in individual molecules. We thus explored the accuracy of the tools for identifying 5mC sites in individual reads. We evaluated the per-read performance of each tool across their range of prediction scores at each site in individual reads, except for DeepMod, which only gives the percentage of methylation per site, so could not be included in this analysis. All tested tools achieved areas under the receiver operating characteristic (ROC) curve (AUC) (Fig. 3a) and areas under the precision-recall (PR) curve AUC_{PR} above 0.8 (Fig. 3b). DeepSignal showed the highest AUC and AUC_{PR} values, closely followed by Nanopolish and Megalodon (Fig. 3b). In contrast, Guppy had decreased precision at high recall values (Fig. 3b). The differing accuracies of methods across different conditions suggest that a consensus approach may capture the advantages of the methods and compensate for the potential pitfalls.

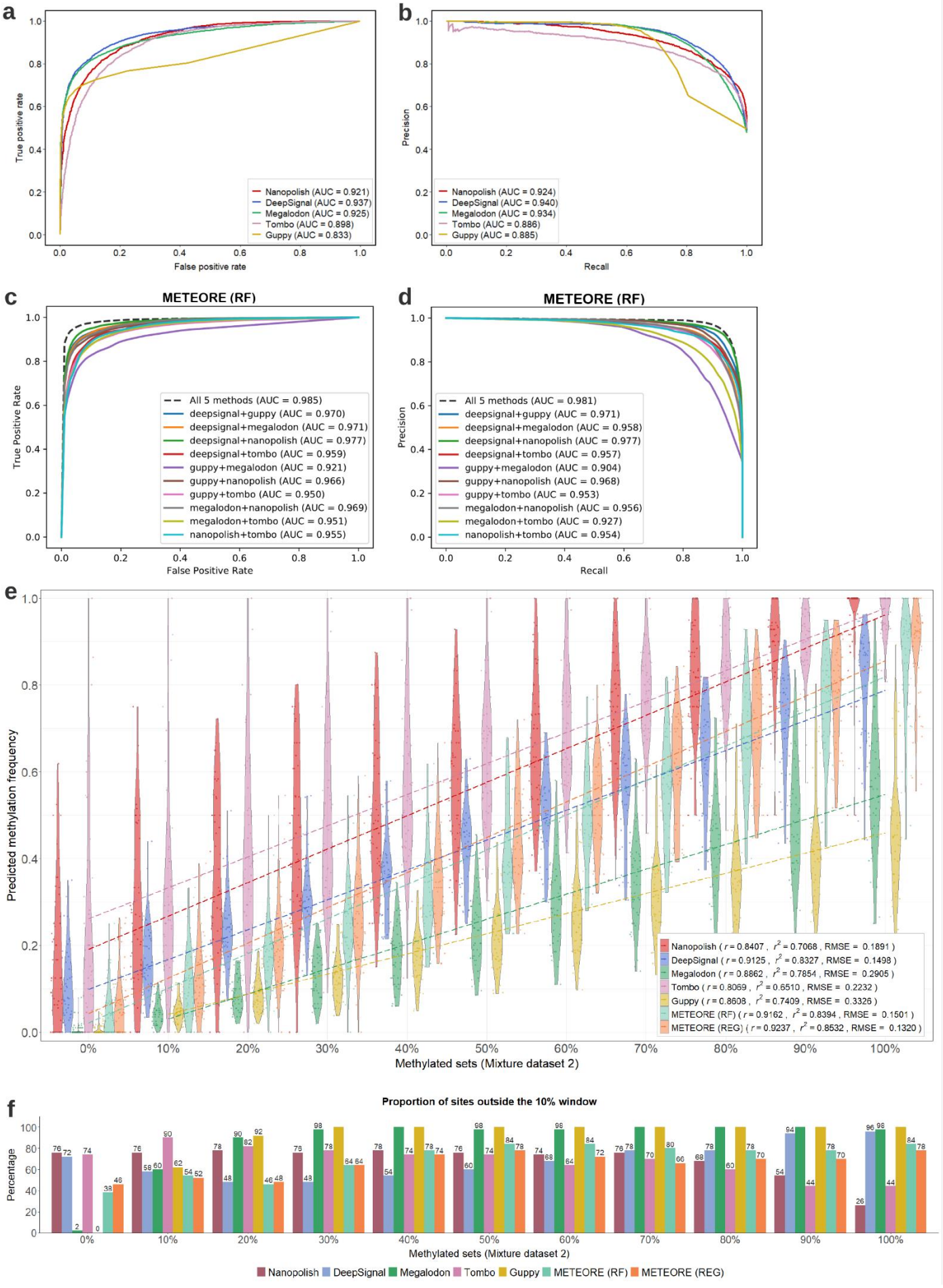

Figure 3. Model accuracy at the individual read level. (a) Receiver operating characteristic (ROC) curves showing the false positive rate (x axis) and true positive rate (y axis) for the predictions at individual reads for the 5 methods tested, using reads from 0% and 100% methylated sets. **(b)** Precision-recall (PR) curves showing the recall (x axis) and precision (y axis) for the predictions at individual read and site levels for the 5 methods tested, using reads from 0% and 100% methylated sets. **(c)** ROC curves for the random forest (RF) models (parameters: max_depth=3 and n_estimator=10) combining two methods, as well as the model combining the 5 methods, built from the average of a 10-fold cross validation with mixture dataset 1. Similar plots for an RF model with default parameters and for a regression model are shown in Supp. Fig. 1. **(d)** PR curves for the same models as in (c). **(e)** Violin plots showing the predicted methylation frequencies (y axis) for each control mixture set with a given proportion of methylated reads (x axis) from the mixture dataset 2 for the five tested tools plus METEORE combining Nanopolish and DeepSignal using random forest (RF) and regression (REG) models. Violin plots showing the predicted methylation frequencies (y axis) for each control mixture set with a given proportion of methylated reads (x axis) for the mixture dataset 2. The Pearson's correlation (r) and coefficient of determination (r^2) are given for each tool. **(f)** We indicate the proportion of sites predicted outside a window around the expected methylation proportion, i.e. a site in the m% dataset was "outside" if the predicted percentage methylation was outside the interval [(m-5)%, (m+5)%] for intermediate methylation values, or outside the intervals [0,5%] or [95%,100%] for the fully unmethylated or fully methylated sets, respectively.

Combination of predictions in individual molecules improve accuracy

We developed a new method, METEORE (https://github.com/comprna/METEORE), to combine two or more prediction tools. We implemented two approaches in METEORE, a random forest (RF) and a linear regression (REG) model. The combination of two methods using either of these two models provided overall an increase in accuracy compared to individual methods (Figs. 3c and 3d) (Supp. Fig. 1). Next, we tested the individual tools together with METEORE (RF and REG) for the combination of DeepSignal and Nanopolish, which showed one of the highest accuracies, in METEORE (RF and REG) together the other five tools on a different collection of methylation mixtures (mixture dataset 2) (Supp. Table 1) (Supp. File 2). The predictions were performed on individual reads, and then summarized per CG site to compare with the expected percentage methylation. METEORE REG combining Nanopolish and DeepSignal achieved higher correlation and lower RMSE compared with the individual tools (Fig. 3e). METEORE RF performed similarly to DeepSignal, and improved accuracy over the other tools (Fig. 3e). METEORE also showed an improvement in the proportion of sites predicted within the expected 10% window at some of the methylation frequency values (Fig. 3f).

Optimized score cutoffs improve the accuracy of methylation predictions

So far we had applied the tools with their default score cutoffs. We reasoned that it might be possible to identify optimal cutoffs for the scores at the individual read level to improve the accuracy of the predictions of methylation frequency per site. To determine these optimal cutoffs, we considered the distribution of scores in individual reads (Supp. Fig. 2a) and several accuracy metrics using mixture dataset 1 (Supp. Fig. 2b). We then determined for each method the score that corresponded to the maximum of TPR – FPR (TPR = true positive rate and FPR = false positive rate) (Supp. Table 3). Applying these scores to separate methylated and unmethylated sites per-read on mixture dataset 2, all tools except Nanopolish improved in the prediction of percentage methylation with respect to the default cutoffs (Supp. Fig. 3a). Selecting the score corresponding to the minimum of $FPR^2 + (1-TPR)^2$ led to similar cutoffs (Supp. Table 3) and improvements in accuracy (Supp. Figs. 3b). Applying optimal cutoffs to METEORE (RF and REG) combining DeepSignal and Nanopolish improved upon the default cutoffs and achieved higher correlation and a lower RMSE compared with the individual tools (Fig. 3e) (Supp. Figs. 3a and 3b). All tools still showed a high proportion of predicted sites outside the expected 10% window. Megalodon, Tombo and Guppy had fewer per-site predictions outside the expected percentage methylation window at the intermediately methylated sites (Supp. Fig. 3c and 3d). METEORE gave a lower proportion of sites predicted outside the expected windows at low methylation frequency (Supp. Figs. 3c and 3d).

Discarding reads of uncertain methylation prediction state improves accuracy

We considered the alternative strategy of discarding sites in reads with uncertain methylation state. This is used by default in Nanopolish and Tombo, which utilize a double cutoff (higher and lower than the point of indecision). To test this strategy, we used the distribution of scores (Supp. Fig. 2) and removed the 10% of sites in individual reads that were closest to the score at which the FPR and 1-TPR curves crossed (Supp. Table 4). Using this approach, all methods, including METEORE, achieved higher correlation and lower RMSE values compared with the default cutoffs (Supp. Fig. 4a). We also considered the scores at which FPR=0.05 and 1-TPR=0.05 and removed all predictions between these two values (Supp. Table 4). This led to improved performance of all methods with respect to default cutoffs, except for Megalodon and Nanopolish (Supp. Fig. 4b). Additionally, we observed that Nanopolish and Megalodon used much fewer reads compared to the other methods (Supp. Fig. 4c).

Nanopore recapitulates bisulfite sequencing data at lowly and highly methylated sites

We next performed a comparison with whole genome bisulfite sequencing (WGBS), which is one of the most used techniques to study CpG methylation at genome-wide scale. To achieve enough read coverage, we used the Cas9-targeted nanopore sequencing (nCATS) protocol[17]. We selected ten regions of forensic relevance to sequence the native nuclear DNA of a human lymphoblastoid cell line (NA12878) with a MinION flowcell (Supp. Table 5) (Supp. Fig. 5). We used the reads corresponding to the targeted regions to analyze the CpG methylation patterns with the tested tools using default cutoffs, and compared the results with existing WGBS data for NA12878 from the ENCODE project[18]. Combining the methylation predictions from both stands on CpG sites showed an improved correlation for all tools compared with using the methylation prediction independently for each strand (Supp. Fig. 6). All tools showed a positive correlation of the Nanopore-based predictions with WGBS signals (Table 1). METEORE (REG) achieved the highest Pearson correlation, closely followed by DeepSignal and METEORE (RF). METEORE (RF and REG) had the lowest RMSE values (Table 1).

	N	***r***	***r²***	***ρ***	***RMSE***
Nanopolish	1704	0.8652	0.7485	0.8326	0.2248
DeepSignal	1731	0.9177	0.8423	0.8765	0.1708
Megalodon	1720	0.8667	0.7512	0.8406	0.2128
Tombo	1661	0.7765	0.6030	0.7537	0.2996
Guppy	1738	0.8513	0.7246	0.8316	0.2334
DeepMod	1739	0.7401	0.5477	0.7264	0.2874
METEORE (RF)	1724	0.9166	0.8402	0.8740	0.1698
METEORE (REG)	1724	0.9178	0.8424	0.8755	0.1682

Table 1. Comparison of CpG methylation frequencies from whole genome bisulfite sequencing (WGBS) Illumina data with Cas9-targeted Nanopore data. For each method we provide the number of sites (N), the Pearson's correlation (r), coefficient of determination (r^2), the Spearman's rank correlation (ρ), and the root mean square error (RMSE) for the comparison of the percentage methylation predicted from Nanopore with the percentage methylation calculated from whole genome bisulfite sequencing (WGBS) data. We show the results for six tested tools and METEORE combining DeepSignal and Nanopolish using a random forest (RF) (parameters: max_depth=3 and n_estimator=10) or a regression (REG) model.

To compare the spread of methylation prediction, we categorized the WGBS data into three bins of increasing methylation frequency (Fig. 4a). Tombo showed the largest dispersion of values at low (0.0-0.3) and intermediate (0.3,0.7) methylation, Guppy and DeepMod underpredicted at high methylation sites (0.7-1.0), and all methods overpredicted at intermediate methylation (Fig. 4a). The same analyses on similar datasets from eight other

regions[17] (Supp. Fig. 7) showed similar trends (Supp. Figs. 8 and 9a). METEORE (RF and REG) together with DeepSignal achieved the highest Pearson correlation and the lowest RMSE values (Supp. Table 7). All methods achieved higher correlations with WGBS data for these regions, possibly because the tested regions contained more CpG sites with high or low methylation (Supp. Table 7).

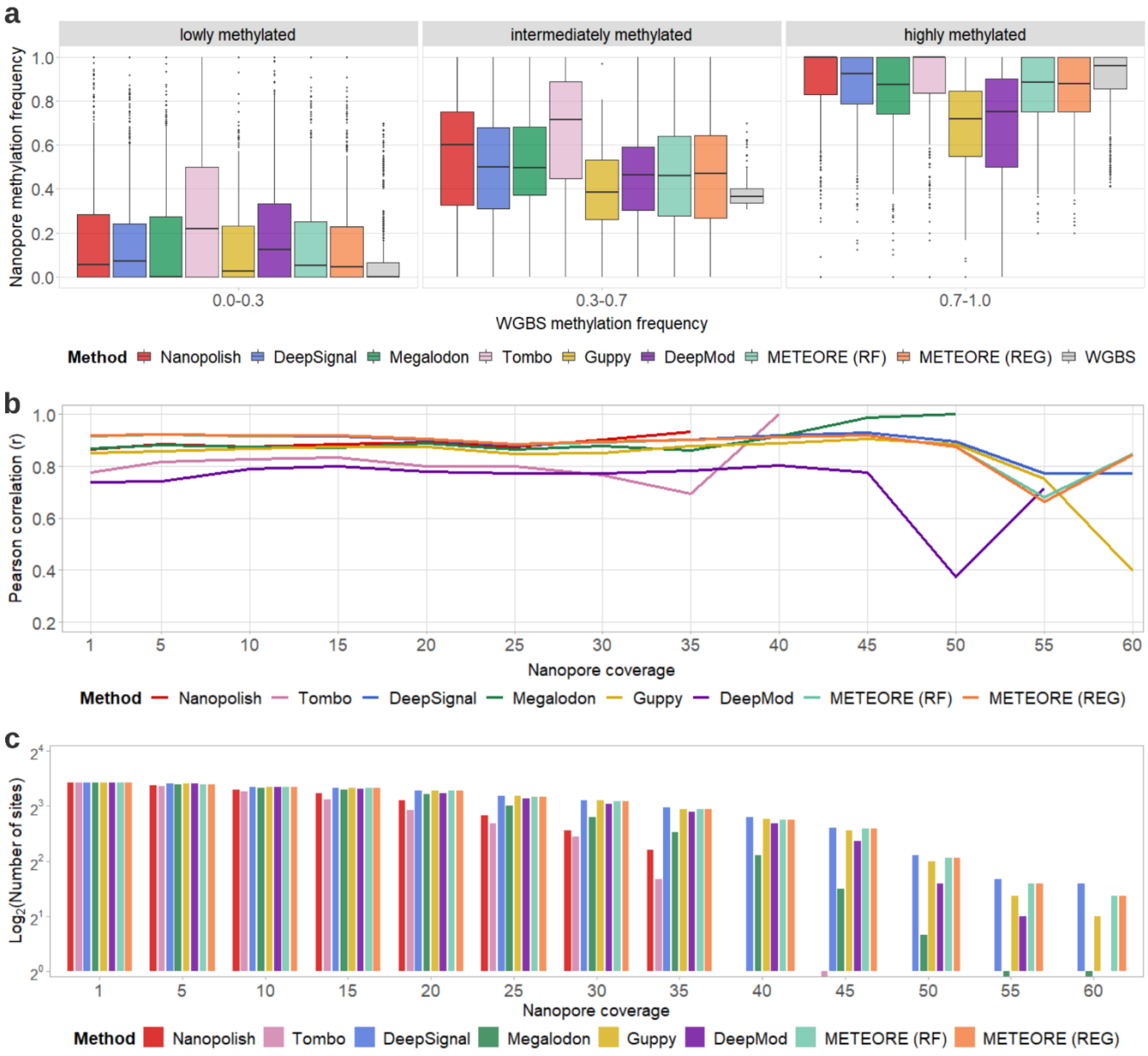

Figure 4. Comparison of CpG methylation predictions from Nanopore with whole genome bisulfite sequencing (WGBS). (a) Distribution of methylation calls from Nanopore (y axis) across three WGBS methylation bins: unmethylated or lowly methylated (0.0-0.3), intermediate methylation (0.3-0.7), and highly or fully methylated (0.7-1.0). **(b)** Pearson's correlation (r) (y axis) between nanopore methylation frequencies calculated from Nanopore by each of the tested tools and WGBS at sites with predictions from both strands combined at each level of minimal coverage, i.e. minimum number of Nanopore reads considered per site (x axis). **(c)** Number of sites on a logarithmic scale (y axis) considered at each value of

minimum coverage in (b). Similar plots for individual sites are provided in Supp. Fig. 10. METEORE (RF) is the combination of DeepSignal and Nanopolish using a random forest (parameters: max_depth=3 and n_estimator=10). METEORE (REG) is the combination of DeepSignal and Nanopolish using a regression model.

Methylation accuracy is stable at low coverage

Overall, correlations between the Nanopore-based predictions and WGBS signals stayed approximately constant at for all tools across different levels of coverage (Fig. 4b) (Supp. Fig. 10). Nanopolish and Megalodon showed a slight improvement at high coverage, but could not predict under high coverage requirements, consistent with them discarding many reads (Fig. 4c). Tombo had lower correlations across all coverage levels, which agreed with our findings that Tombo was less accurate in separating fully methylated from fully unmethylated sites. The same analyses with data from Gilpatrick et al.[17] also showed stable correlations with WGBS at most levels of coverage for all tools, with a marked drop at high minimum coverage (Supp. Figs. 9b and 9c). As with our nCATS dataset, Nanopolish achieved the highest correlation values but ran out of sites at high minimum coverage (Supp. Fig. 9).

METEORE recovers the methylation patterns along genomic regions

We compared the methylation profiles from WGBS and Nanopore along our ten tested regions (Fig. 5) (Supp. Fig. 11). In general, Guppy and DeepMod tended to underpredict relative to WGBS, whereas Tombo tended to overpredict. DeepSignal and METEORE (RF and REG) were consistent with the overall WGBS pattern independently of coverage and CG content (Fig. 5). All tools recapitulated the known pattern of hypomethylation at CpG islands (CGIs)[19] (Supp. Figs. 11). However, there were local inconsistencies by some of the tools. For instance, in the region spanning the first and second introns of *IRF4* (chr6:392228-401463) (Fig. 5a), Guppy and DeepMod underpredicted the methylation frequency. In the region spanning the genes *ACKR1* and *CADM3* (chr1:159199780-159212236), Tombo overpredicted the methylation frequency, and Guppy failed to predict an increase in methylation described by WGBS and the other tools (Fig. 5b). At the *TPO* locus (Fig. 5c), the intermediate methylation at a CGI described by WGBS was recovered by all tools, except Tombo and Nanopolish, which overpredicted the methylation frequency. Using the eight different regions from Gilpatrick et al.[17] we found similar results (Supp. Fig. 12). For instance, Tombo overpredicted the methylation frequency at a CGI in the *GPX1* promoter (chr3: 49352525-49366169) (Fig. 5d).

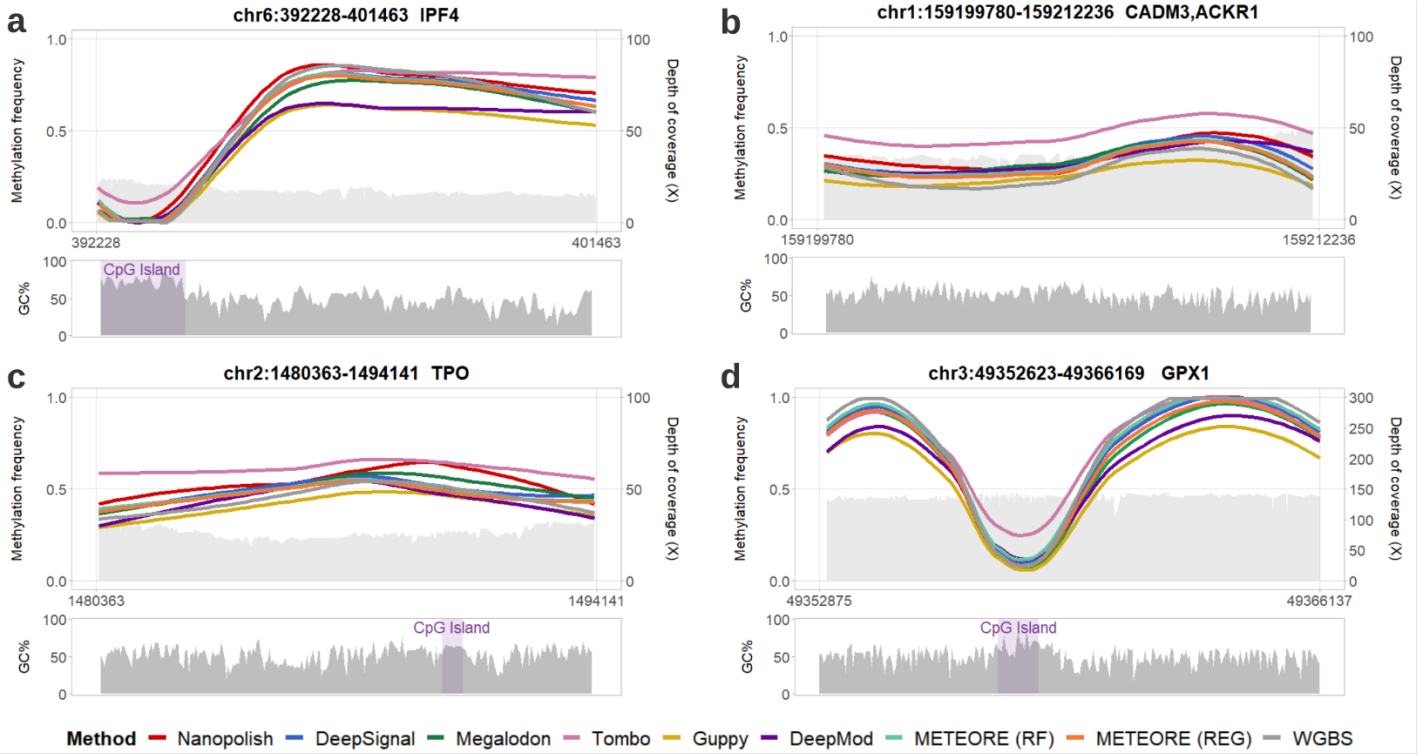

Figure 5. Comparison of CpG methylation predictions from Nanopore with whole genome bisulfite sequencing (WGBS) along Cas9-targeted regions. LOESS smoothing line plots of methylation calls frequency predictions (left y axes) from WGBS Illumina and from Nanopore data using seven tools: Nanopolish, DeepSignal, Megalodon, Tombo, Guppy, DeepMod, and METEORE random forest (RF) and regression (REG) models. The plots include the Nanopore read coverage (right y axes), shown as a grey area. The panels below show the GC content of the region, using a window size of 50 bases. The depicted regions are **(a)** chr6:392228-401463, which covers the first and second introns of gene *IRF4*; **(b)** chr1:159199780-159212236, which covers the genes *ACKR1* and *CADM3;* **(c)** chr2:1480363-1494141, which covers the gene *TPO*; **(d)** chr3:49352525-49366169, which covers the genes *GPX1* and *RHOA.*

Finally, we used the optimized thresholds derived before from the individual read analysis for the different methods (Supp. Table 3). Using a single score cutoff based on the maximum value of TPR-FPR, we observed an improvement in DeepSignal and Tombo, but similar or slight worse accuracies in the other methods (Supp. Table 8). Using the strategy of a double cutoff based on the removal of the 10% of reads around the crossover point between methylated and unmethylated scores (Supp. Table 4), the concordance with WGBS improved only for Nanopolish, DeepSignal, Tombo and METEORE (REG) (Supp. Table 9).

Discussion

Our systematic benchmarking of DNA methylation prediction from Nanopore sequencing using individual reads, controlled methylation mixtures, Cas9-targeted sequencing, and whole genome bisulfite sequencing, indicated that no single method predicts correctly across all ranges of methylation frequency. Extreme cases were Guppy and Megalodon, which correctly identified unmethylated sites but failed at fully methylated sites, as well as Nanopolish and Tombo, which were able to recover fully methylated sites but had a high rate of false positives at unmethylated sites. Moreover, the predictions per-site generally showed a high dispersion and did not generally agree with the expected methylation frequency. These issues motivated us to propose a new consensus method, METEORE, aiming to ameliorate the errors from some methods and incorporate the advantages from others. The combination of two methods improved the overall accuracy at the levels of individual reads and per-site methylation frequency. Our analyses suggested that it is generally advantageous to run at least two tools to obtain an accurate picture of the DNA methylation patterns. Although combining five tools improves accuracy even further, it might be impractical for routine analyses, due to the running times for some methods without GPU support (Supp. Table 10). The combination of Nanopolish and DeepSignal struck a good balance for accuracy and running times. Nanopolish was faster, but DeepSignal achieved generally better accuracy. Although tools like Guppy and Megalodon were much faster when running with GPUs, the overall accuracy was not as good as for the other methods.

Additionally, we found that by reassessing the score cutoffs for individual reads, the per-site methylation predictions could be improved. This suggests that DNA methylation prediction is performed suboptimally for most methods. Related to this, we observed that there was an advantage in removing sites with uncertain methylation status. This strategy improved the accuracy for most methods and did not have a large impact on the number of reads, except for Nanopolish and Megalodon, where the number of reads available was significantly reduced.

The comparison with whole genome bisulfite sequencing (WGBS) datasets using two independent experiments of Cas9-targeted Nanopore sequencing recovered results consistent with the analyses with individual reads and control mixture datasets. Although the tools recovered the overall WGBS patterns across different genomic regions, and independently of GC content and coverage, there were some remarkable variations. In particular, Nanopolish tended to overpredict methylation values, DeepMod, Guppy and Megalodon tended to underpredict, and Tombo and Guppy showed local discrepancies with the other tools. DeepSignal and METEORE achieved overall good consistency with the WGBS data. Furthermore, we observed limited impact of coverage on accuracy. This suggests a potential for the development of sensitive diagnostic and forensic tests without relying on high coverage requirements. Moreover, the use of our improved strategies with optimized cutoffs or with METEORE

consensus predictions will facilitate the development of analyses based on individual reads. In summary, we highlighted the strengths and weakness of state-of-the-art methods to predict DNA methylation from Nanopore sequencing and provided various new strategies to improve the prediction accuracy. We expect METEORE and the provided pipelines this will facilitate the accurate analysis of genome-wide methylation patterns both per site and in individual molecules in multiple biological contexts.

Methods

Data availability

Nanopore sequencing data generated in this study has been deposited in the Sequence Read Archive (SRA) under study accession PRJNA656260 (https://www.ncbi.nlm.nih.gov/sra/PRJNA656260). Nanopore sequencing data from Gilpatrick et al.[17] used in this study is available in SRA under study accession PRJNA531320 (https://www.ncbi.nlm.nih.gov/sra/PRJNA531320). Nanopore sequencing data for *E. coli* methylated and unmethylated genomes used in this study is available at the European Nucleotide Archive (ENA) study accession ERP014559 (https://www.ebi.ac.uk/ena/browser/view/PRJEB13021)[8]. WGBS data from the ENCODE project[18] used in this study are available at https://www.encodeproject.org/ under IDs ENCFF279HCL and ENCFF835NTC.

Software availability

Software developed in this study is available at https://github.com/comprna/METEORE under the MIT license.

DNA methylation of controlled mixtures

We used *E. coli* (K12 ER2925) control reads[8], for which DNA was amplified by PCR (unmethylated, negative control) and half of the PCR-amplified DNA was subjected to CpG methyltransferase (M.SssI) treatment to methylate cytosines at CpG sites (methylated, positive control). It was reported that M.SSsI has an efficiency of 95%[8]. This would not affect the benchmarking with the fully unmethylated set (0% methylated) but it might affect the other mixture datasets. Accordingly, to avoid potential biases in our benchmarking analysis, we considered bins of percentage methylation ranges, in steps of 10%. For instance, the bin of the fully methylated sites included sites of 90-100% methylation. After alignment of all PASS reads for both controls using minimap2[20] and removing secondary and supplementary reads, 110,795 reads of unmethylated control and 69,453 reads of methylated control remained. From the 346,793 CG sites in the E. coli reference genome (NC_000913.3), we selected 100 random sites with a single CpG in a 20nt window (NNNNNNNNNCGNNNNNNNNN, with no CG in the region with Ns) that had aligned sequencing reads from both positive and negative control datasets. Using the filtered PASS reads covering these 100 selected CpG sites from the control datasets, we created 11

benchmarking datasets with specific mixtures of methylated and unmethylated reads, namely, containing 0%, 10%,…, 90%, 100% of methylated reads, which we used for the initial benchmarking of the tools to predict different methylation mixtures per genomic site (mixture dataset 1). For independent validation, we built an independent mixture dataset 2. We selected a different set of 50 CG sites (single CG in a 20nt window), and further selected 6,803 different reads from the remaining fully methylated or fully methylated reads not used in the dataset above. We built again 11 benchmarking datasets with specific mixtures of methylated and unmethylated reads, namely, containing 0%, 10%, 20%, …, 90%, 100% of methylated reads (mixture dataset 2). We used this dataset to obtain the correlation of the different methods with the expected percentage methylations using the optimal thresholds calculated before. We also used this dataset to test our combined model (see below).

CpG methylation detection with six different tools

We developed a standardized workflow for six tools: Nanopolish[8], Megalodon[11], DeepSignal[12], Guppy[13], Tombo[14] and DeepMod[15]. Snakemake pipelines to run these tools are available at (https://github.com/comprna/METEORE). When we selected the tools to be included in this study, we excluded NanoMod[21], as it only allowed detection of methylation differences using a control and a test sample; SignalAlign[10], as its repository had not been updated for over four years; and mCaller[22], because it only had been trained for 6mA, but not for 5mC.

Methylation level was collected for individual CpG sites in both strands and considered per site and per read or summarized (methylation frequency) per site. For the comparison with bi-sulfite sequencing, we also considered the approach of merging the CpG methylation calls from both strands into a single strand. This was done in the following way: For the sites that had methylation frequencies from both strands, we combined the sites by averaging the methylation frequencies and adding up the coverage. Otherwise, we kept the methylation frequency the same if the site only had methylation prediction from one strand.

Nanopolish (v0.13.2) assigns a log-likelihood ratio to each individual CpG site or to a group of nearby CpG sites that share the same methylation level in each site within the group. A positive log-likelihood ratio value indicates evidence of methylation. To include all the predictions per read and per site, such CpG groups were split up into the constituent sites with the same log-likelihood ratio using a Python script incorporated in our Snakemake pipeline (https://github.com/comprna/METEORE). The output file was further processed in R. As default cutoffs we considered those suggest by Simpson et al.[8], i.e. log-likelihood >2.5 for methylated sites and <-2.5 for unmethylated sites. The methylation frequency was then calculated for each site as the number of mapped reads predicted as methylated divided by the number of total mapped reads. For the benchmarking analysis for sites in individual reads and to establish the optimal cutoffs we used the log-likelihood score given by Nanopolish.

Tombo (v1.5.1) implements three approaches to detect nucleotide modifications: 1) the alternative base detection approach, which computes a statistics by scaling log likelihood ratios to identify targeted bases where the signals match the expected level for a non-canonical base; 2) the *de novo* approach, which performs a hypothesis test by statistically comparing signals to an in-silico reference; and 3) the sample comparison approach. This latter approach provides two different ways for modified base detection, one uses a canonical model adjusted by a control set of reads to identify deviations between expected and observed levels, while the other one compares signal levels from two sets of reads at each reference position. Of the three approaches, we used the alternative model specific for CpG to be able to predict CpG methylation in individual samples and reads. This model tests the signal levels against expected canonical and alternate 5mC in at CG motifs, producing the per-read binary statistics in HDF5 format, where positive values indicate canonical bases and negative values modified bases. For the initial analysis, we used the default cutoffs of (-1.5, 2.5) where scores below -1.5 were considered as methylated and above 2.5 unmethylated, and scores between these thresholds did not contribute to the per-site methylation. Tombo outputs four individual wiggle files (one per strand) reporting the read coverage level and the raw fraction of 5mC at genome level. These files were converted to TSV format with a Python script included in our Snakemake pipeline, with the score and coverage for each mapped CpG site for downstream analyses. For the benchmarking analysis per site and per read, we used the per-read binary statistics given by Tombo.

DeepSignal (v0.1.7) predicts the methylation state of the targeted cytosine at CpG motifs and outputs the probabilities of being methylated, P(m), and unmethylated, P(u), for each cytosine in each read. For the initial analysis, a base was considered as methylated if the methylated probability was greater than the unmethylated probability, P(m)>P(u), and unmethylated otherwise. Additionally, a Python script was added to our Snakemake pipeline to calculate the methylation frequency at each CpG site. For the benchmarking analyses per site and per read, we used a score calculated as the log2 ratio of the probabilities, i.e. $\log_2(P(m)/P(u))$.

Guppy (v3.6.0) is only available to members of the Nanopore community (https://community.nanoporetech.com). Guppy basecalls 5mC at CG sites as a fifth base along with the four canonical DNA bases. We used the configuration file named dna_r9.4.1_450bps_modbases_dam-dcm-cpg_hac.cfg when to run running Guppy's modified basecalling model. This produced the reads supporting the modifications in fast5 format. To process and analyse the Guppy's fast5 output, we used the pipeline provided at https://github.com/kpalin/gcf52ref, which was incorporated into our Snakemake pipeline (https://github.com/comprna/METEORE).

Megalodon (v2.1.0) uses Guppy for modified basecalling and identifies 5mC by anchoring the basecalling output to the reference and assigning a score for the candidate modified base. Megalodon requires Guppy to be installed and the path to Guppy basecalling executable server to be set. Megalodon produces per-read modified base log probability at each mapped CpG site. A default threshold of 0.8 was used as a minimum score to include a

modified basecall in the final aggregated output in the bedMethyl format file containing per-site coverage and methylation percentage. For the benchmarking analysis per site and per read, we used a log-likelihood score calculated by subtracting the natural log-probability of the modified base, log(M) and the natural log-probability of the canonical base, log(C), resulting in log(M/C).

DeepMod takes single-read fast5 files as input and provides as output a methylation prediction summary per site at genome level for each strand in a BED format containing the coverage, number of methylated reads and methylation percentage. As DeepMod did not provide any information for individual reads, we could not use it for the per-read benchmarking analyses.

METEORE consensus models

METEORE was created to provide a consensus prediction using the scores from two or more methods. As the two main approaches for supervised learning are classification and regression, we implemented both types of methods in METEORE to test the combination of tools. For the classification model, we used a random forest (RF) classifier[23] from the Python sklearn library[24]. We scaled the prediction scores from each individual method to the range of [0,1] using min-max scaling[23.] The Receiver operating characteristic (ROC) and precision-recall (PR) curves were built from a 10-fold cross-validation on the prediction scores from the input methods to produce receiver operating characteristic (ROC) and precision-recall (PR) curves. The reads were randomly selected during cross-fold validation. For the METEORE implementation we used the parameters max_dep=3 and n_estimator=10. We also tested tested the default parameters of RF from sklearn (n_estimator = 100 and max_dep = None). The METEORE RF model tested in our analyses was trained on the entire mixture dataset 1 with the scores from Nanopolish and DeepSignal and using parameters max_dep=3 and n_estimator=10. The regression-based approach used sklearn's RidgeCV linear regression (REG) model, and min-max scaling as in the RF model. The initial ROC and PR curves were produced with the built-in 5-fold cross validation in RidgeCV. The METEORE REG model tested in our analyses was trained on the entire mixture dataset 1 with the scores from Nanopolish and DeepSignal. After prediction, resulting scores were classified as unmethylated if less than 0.5 and methylated if greater. METEORE RF and REG only considered reads for which there were prediction scores from both the combined input methods. The scripts to train and run METEORE are available at https://github.com/comprna/METEORE.

Processing of per-site methylation calls

The raw output of each of the tools contained methylation information for each aggregated CpG site on both strands. That is, each mapped CpG site on the positive strand of the human reference genome had a counterpart CpG site mapped on the negative strand. To perform the per-site benchmarking, we used the positive strand coordinate system, so the mapped sites that were on the negative strand were lifted to positive coordinates by

subtracting 1 from the coordinate position. If there were predictions made on both strands for the same site, we obtained the mean of methylation frequency for that site. If there was no per-site information only for one of the strands, we kept the same prediction from that strand, and lifted it to the positive strand if necessary.

Guide RNA design and RNP complex assembly

We used the Cas9-targeted nanopore sequencing (nCATS) protocol[17] to target ten regions of the human genome. This PCR-free protocol uses Cas9 to cut double-stranded DNAs (dsDNAs) at specific sites and then preferentially ligates sequencing adapters to the cleaved ends for enrichment[17]. The cleaved target DNA strands with adapters attached are then sequenced. We designed ten pairs of RNA CRISPR guides (crRNAs) for ten forensically relevant regions (Supp. Table 5). An initial panel of candidate crRNAs was designed using the freely available tool CHOPCHOP[25] as recommended in the ONT protocol. For each target region, about three to five best crRNAs were selected based on the cleavage location, crRNA efficiency, and the number of predicted mismatches using CHOPCHOP. The crRNAs were then evaluated by the IDT's design checker[26] to select for high on-target performance and low off-target activity. Ten pairs of guide RNA were used to enrich ten human regions ranging from 8-36kb. Here, one gRNA was used on either side of each target region (Supp. Table 6). Each RNA oligo including crRNAs (IDT, custom designed) and tracrRNA (IDT, 11-05-01-12) was resuspended in IDTE buffer pH 7.5 (IDT, 11-01-02-02) to a final concentration of 100uM. crRNAs were then pooled to make an equimolar crRNA mix by combining equal volumes of each crRNA. In order to form gRNA duplexes, the crRNA pool and tracrRNA were then combined in equimolar concentrations with duplex buffer (IDT, 11-05-01-12), followed by denaturation for 5 min at 95 °C, then allowing to cool to room temperature. RNP complexes were created by assembling the following components: gRNA duplexes, CutSmart buffer (NEB, B7204), nuclease-free water and Cas9 endonuclease (IDT, 1081058).

Library Preparation

Genomic DNA (gDNA) from the GM12878 human cell line was obtained from the Coriell Institute (coriell.org) (cat. No. NA12878). The purity of the purchased gDNA was measured with the Nanodrop spectrophotometer (Thermo Fisher) at the 260/280nm and 260/230 nm values. Dephosphorylation of 5' ends of DNA was performed as follows. A 5ug amount of input DNA was dephosphorylated to prevent downstream adapter ligation using Quick dephosphorylation Kit (NEB, M0508). DNAs were resuspended in the CutSmart buffer and dephosphorylated with Quick CIP enzyme in a PCR tube for 10 min at 37 °C, followed by heating for 2 min at 80 °C for CIP enzyme inactivation. Cas9 Cleavage and dA-tailing was performed as follows. A 100 mM aliquot of dATP (NEB, N0440S) was first diluted to 10 mM. After allowing the dephosphorylated DNA sample to return to room temperature, the pre-assembled RNP complexes, 10 mM dATP and Taq DNA Polymerase with Standard Taq Buffer (NEB, M0273) were added to the PCR tube containing the sample for the *in vitro* digestion reaction and subsequent dA-tailing. The sample was then incubated at 37 °C for 15 min, and then at 72 °C for 5 min. By

dephosphorylating pre-existing DNA ends prior to Cas9 cleavage, sequencing adapters and ligation buffer from the Oxford Nanopore Ligation Sequencing Kit (ONT, LSK109) were preferentially ligated to the cleaved DNA ends at Cas9 cleavage sites using T4 Ligase from the NEBNext Quick Ligation Module (NEB, E6056) for 10 min at room temperature. The sample was cleaned up to remove excess adapters using the Agencourt AMPure XP beads (Beckman Coulter, A63881), washing twice on a magnetic rack with the long-fragment buffer (ONT, LSK109) before eluting in 14 µl of elution buffer (ONT, LSK109). A 1 µl aliquot of the final library was quantified using the Qubit dsDNA Broad Range assay kit (Thermo Fisher). Starting with 5 µg of input DNA, 1 µg of DNA was recovered after library preparation. The library was stored on ice until ready to load.

Sequencing and data processing

Before loading, the flowcell was primed with a solution consisting of flush buffer (ONT, LSK109) and flush tether (ONT, LSK109). The sequencing library was prepared by adding sequencing buffer (ONT, LSK109) and loading beads (ONT, LSK109) into the DNA library, and then loaded into the flowcell. The sample was run on a MinION flow cell (FLO-MIN106, R9.4.1 pore) using the MinION sequencer for 19 hr, operated using the MinKNOW software. Live basecalling was carried out during the experiment using Guppy's fast basecalling model in MinKNOW. The resulted FASTQ files were immediately aligned to the human reference genome (GRCh38/hg38) using minimap2[20], followed by visualization with IGV to confirm generation of on-target sequencing reads. Post-run basecalling was performed using Guppy (v.3.2.4) high-accuracy model to generate the final set of sequencing reads with higher read accuracy than the fast model and recognition of modified bases from the electrical signal data. Reads were aligned to the human reference genome (GRCh38/hg38) using minimap2. Using Samtools, we collected aligned reads within the enriched regions as on-target reads. Those outside the targeted regions were considered off-target reads and subsequently discarded.

Comparison with bisulfite sequencing data

We compared with the published whole genome bi-sulfite sequencing (WGBS) data for NA12878 (Encyclopedia of DNA Elements (ENCODE) accessions: ENCFF279HCL, ENCFF835NTC) using two different approaches. In the first approach, we compared every CpG site on both strands for nanopore and WGBS data, preserving the strand information. In the second approach, we used the positive strand coordinate system by lifting all sites from the negative stand to be on the positive strand, i.e. site position – 1. For the sites that had methylation evidence from both strands, we combined the sites by averaging the methylation frequencies and adding up the coverage. We preserved the information if the site only had methylation prediction from the positive strand or the negative strand.

To obtain high confidence methylation calls from WGBS data for validation, the resulting individual or combined CpG sites were processed in the following way. CpG sites with zero coverage from both WGBS replicates were

discarded. Furthermore, we calculated the difference in methylation frequency between both WGBS replicates and considered the 0.1 and 0.9 quantiles of the distribution of differences. A CpG site was kept if the difference for that site was between those 0.1 and 0.9 quantiles, otherwise it was removed. We finally calculated the Pearson and Spearman correlations between methylation frequencies calculated from WGBS and those calculated from Nanopore reads by each of the tested methods using R. We also subsampled the reads for the CpG sites that were covered with at least 1, 5, 10, 15, 20, etc. reads and calculated Pearson correlation at different levels of read coverage for further evaluation.

References

1. Greenberg, M.V.C. and D. Bourc'his, *The diverse roles of DNA methylation in mammalian development and disease.* Nature Reviews Molecular Cell Biology, 2019. **20**(10): p. 590-607.

2. Kader, F. and M. Ghai, *DNA methylation and application in forensic sciences.* Forensic science international, 2015. **249**: p. 255-265.

3. Jones, P.A., *Functions of DNA methylation: islands, start sites, gene bodies and beyond.* Nature Reviews Genetics, 2012. **13**(7): p. 484-492.

4. Yong, W.-S., F.-M. Hsu, and P.-Y. Chen, *Profiling genome-wide DNA methylation.* Epigenetics & Chromatin, 2016. **9**(1): p. 26.

5. Raiber, E.-A., R. Hardisty, P. van Delft, and S. Balasubramanian, *Mapping and elucidating the function of modified bases in DNA.* Nature Reviews Chemistry, 2017. **1**(9): p. 0069.

6. Grunau, C., S. Clark, and A. Rosenthal, *Bisulfite genomic sequencing: systematic investigation of critical experimental parameters.* Nucleic acids research, 2001. **29**(13): p. e65-e65.

7. Ehrich, M., S. Zoll, S. Sur, and D. Van Den Boom, *A new method for accurate assessment of DNA quality after bisulfite treatment.* Nucleic acids research, 2007. **35**(5): p. e29.

8. Simpson, J.T., R.E. Workman, P.C. Zuzarte, M. David, L.J. Dursi, et al., *Detecting DNA cytosine methylation using nanopore sequencing.* Nature Methods, 2017. **14**(4): p. 407-410.

9. Laszlo, A.H., I.M. Derrington, H. Brinkerhoff, K.W. Langford, I.C. Nova, et al., *Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA.* Proceedings of the National Academy of Sciences, 2013. **110**(47): p. 18904-18909.

10. Rand, A.C., M. Jain, J.M. Eizenga, A. Musselman-Brown, H.E. Olsen, et al., *Mapping DNA methylation with high-throughput nanopore sequencing.* Nature Methods, 2017. **14**(4): p. 411-413.

11. Oxford Nanopore Technologies. *Oxford Nanopore Technologies GitHub - Megalodon* 2020 [cited 2020 30 June]; Available from: https://github.com/nanoporetech/megalodon.

12. Ni, P., N. Huang, Z. Zhang, D.-P. Wang, F. Liang, et al., *DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning.* Bioinformatics, 2019. **35**(22): p. 4586-4595.

13. Oxford Nanopore Technologies. *Oxford Nanopore Technologies GitHub*. 2020 [cited 2020 25 Apr]; Available from: https://github.com/nanoporetech.

14. Stoiber, M., J. Quick, R. Egan, J. Eun Lee, S. Celniker, et al., *De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing.* bioRxiv, 2017: p. 094672.

15. Liu, Q., L. Fang, G. Yu, D. Wang, C.-L. Xiao, et al., *Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data.* Nature Communications, 2019. **10**(1): p. 2449.

16. Köster, J. and S. Rahmann, *Snakemake—a scalable bioinformatics workflow engine.* Bioinformatics, 2012. **28**(19): p. 2520-2522.

17. Gilpatrick, T., I. Lee, J.E. Graham, E. Raimondeau, R. Bowen, et al., *Targeted nanopore sequencing with Cas9-guided adapter ligation.* Nature Biotechnology, 2020.

18. Dunham, I., A. Kundaje, S.F. Aldred, P.J. Collins, C.A. Davis, et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

19. Chen, P.-Y., S. Feng, J.W.J. Joo, S.E. Jacobsen, and M. Pellegrini, *A comparative analysis of DNA methylation across human embryonic stem cell lines.* Genome Biology, 2011. **12**(7): p. R62.

20. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.

21. Liu, Q., D.C. Georgieva, D. Egli, and K. Wang, *NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data.* BMC Genomics, 2019. **20**(1): p. 78.

22. McIntyre, A.B.R., N. Alexander, K. Grigorev, D. Bezdan, H. Sichtig, et al., *Single-molecule sequencing detection of N6-methyladenine in microbial reference materials.* Nature Communications, 2019. **10**(1): p. 579.

23. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

24. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.

25. Labun, K., T.G. Montague, M. Krause, Y.N. Torres Cleuren, H. Tjeldnes, et al., *CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing.* Nucleic Acids Research, 2019. **47**(W1): p. W171-W174.

26. Integrated DNA Technologies. *CRISPR-Cas9 guide RNA design checker*. 2019; Available from: https://sg.idtdna.com/site/order/designtool/index/CRISPR_SEQUENCE.