Comprehensive benchmarking of tools for nanopore-based detection of DNA methylation

Onkar Kulkarni[1,2,*], Reuben Jacob Mathew[1,*], Lamuk Zaveri[1,*], Rhea Jana[1], Nitesh Kumar Singh[1], Sreenivas Ara[1], Karthik Bharadwaj Tallapaka[1,2], Divya Tej Sowpati[1,2,†]

1 - CSIR Centre for Cellular and Molecular Biology, Hyderabad, India
2 - Academy of Scientific and Innovative Research, Ghaziabad, India

* - These authors contributed equally
† - Corresponding author

Address for correspondence:

Divya Tej Sowpati
CSIR - Centre for Cellular and Molecular Biology
Uppal Road, Habsiguda, Hyderabad - 500007, India
Email: tej@ccmb.res.in
Ph: +91-40-27192862

Abstract

Long read sequencing technologies such as Oxford Nanopore (ONT) offer direct detection of DNA base modifications. While several tools and models have been developed to identify DNA methylation from nanopore data, they are often restricted to 5-methylcytosine (5mC) and older flowcell (FC) chemistry. The performance and accuracy of newer models, including those developed by ONT, remain unknown, particularly for their new FC chemistry (R10.4.1) and sampling rate (5kHz). Here, using a variety of bacterial and human datasets, we systematically evaluate the performance of existing methylation models for 5mC (both CpG and non-CpG contexts), 6-methyladenine, and 4-methylcytosine. We also demonstrate the effect of other parameters such as sequencing depth, read quality, basecalling mode, and more importantly, the presence of neighboring DNA modifications. Our work thus provides important information to researchers utilizing nanopore sequencing to study DNA modifications, and highlights the lacunae in current generation methylation detection models.

Keywords: Nanopore, methylation, 5-methylcytosine, 5mC, 6-methyladenine, 6mA, epigenomics

Introduction

DNA base modifications are the addition of chemical groups to the canonical bases of DNA. Base modifications such as methylation contribute to epigenetic control of gene expression, genome stability, and genome organization[1,2]. The dynamic nature of DNA methylation has been shown to play crucial roles in development, stress, and aging[3–5]. Depending on the canonical base and the position in the carbon ring to which the methyl group is added, there exist several distinct modified DNA bases, the most common being 5-methylcytosine (5mC - addition of methyl group to the 5th carbon of cytosine), 6-methyladenine (6mA - methyl group at the 6th carbon of adenine), and 4-methylcytosine (4mC - addition of methyl group to the 4th carbon of cytosine)[6]. 5mC, specifically in a CG dinucleotide context (CpG) is predominantly found in mammals and its presence in promoters is associated with gene silencing[7]. 6mA and 4mC are more commonly found in prokaryotes, where they mainly contribute to genome protection via restriction modification systems and methylation-mediated mismatch repair[8,9]. However, recent studies report the presence of 6mA and 4mC in eukaryotes as well, and in roles of gene regulation[10,11].

Traditionally, 5mC has been studied using conversion-based methods such as treatment with sodium bisulfite followed by high-throughput short-read sequencing[12]. In addition to being damaging to DNA, bisulfite sequencing is prone to amplification biases as well as difficulty in mapping 5mC in challenging genomic regions[13,14]. Further, other marks such as 6mA and 4mC are not amenable to bisulfite sequencing - their genome-wide analyses have relied on antibody-based methods which lack nucleotide level resolution[15]. Third-generation, single molecule long read sequencing methods such as PacBio and Oxford Nanopore Technologies (ONT) offer simultaneous detection of DNA base modifications from sequencing data.

Nanopore sequencing works by monitoring electrical fluctuations as a DNA strand is translocated through a biological pore[16]. The electrical signal is converted to the corresponding DNA sequence using a basecaller. ONT have constantly upgraded their nanopores for better accuracy and stability. Their older flowcells (FCs) with R9 chemistry used the *Escherichia coli* Curli sigma S-dependent growth (CsgG) protein as the channel, and achieved a modal accuracy of ~95%. The current generation FCs with R10 nanopores use a bioengineered CsgG-CsgF heterodimer. In addition to being more stable, this “double-barrel” nanopore achieves a modal accuracy of >99%, owing to its longer sensing region, and improves accuracy in homopolymer stretches of up to 9nt[17]. ONT has also recently changed its sampling rate from 4kHz to 5kHz, and demonstrated improved sequencing accuracy with the higher sampling rate[18]. In parallel, there have been steady improvements to the basecallers. Their latest libtorch based tool Dorado offers significant improvements in accuracy and is optimized for utilizing GPU for speed[19]. As recently as May 2024, ONT released their v5 models for basecalling which use transformer models as opposed to the traditional LSTM based neural network models[20]. The new models claim a raw-read modal q-score of 26, and improved accuracy for identification of DNA base modifications when tested on synthetic controls[21]. However, the performance of various nanopore models on real data has not been documented so far. In addition to the models offered by ONT, two other tools support identification of 5mC in CpG context: DeepMod2

and RockFish[22,23]. Previous benchmarking studies to understand the accuracy of base modification detection from nanopore were restricted to the older R9.4.1 chemistry, or only in the context of CpG methylation[24–27]. Therefore, the efficacy of nanopore sequencing in studying 5mC in non-CpG context, 6mA, and 4mC remains unknown.

Here, using different bacterial genome datasets as well as the widely studied Genome-In-A-Bottle HG002 data, we benchmark the performance of the base modification models available for current generation nanopore data. In addition to understanding the accuracy of these models in identifying various modified bases, we also evaluate the effect of sequencing depth, quality, and the presence of neighboring base modifications.

Results

Benchmarking strategy

Most existing studies have focussed on benchmarking 5-methylcytosine (5mC), more specifically in CpG dinucleotide context. We set out to perform a comprehensive evaluation of models beyond CpG methylation. The overall strategy, including the datasets generated and used for evaluation, as well as the pipelines followed, is outlined in Figure 1. To generate benchmark datasets for other contexts of 5mC and other DNA modifications, we used both human and bacterial data. For human data, we relied on the extensively cataloged HG002 (GM24385) sample from the Genome-In-A-Bottle (GIAB) consortium, for which the ground truth bisulfite data was available in the public domain. We sequenced the genomic DNA (gDNA) of HG002 using both 4kHz and 5kHz sampling rates, targeting a depth of >60x.

For bacterial data, we mainly focussed on two species: *Escherichia coli* str. K12 substr. MG1655 (*E.coli*) and *Helicobacter pylori* str. 26695 (*H.pylori*). Bacterial methylation systems are highly-motif driven - methylation is found in specific sequence contexts, and often at close to 100% methylation levels[28]. In addition to the wild type (WT) DNA, we had generated data which was devoid of methylation. For *E.coli*, we created a mutant strain where both the major methylases, *dam* and *dcm*, were knocked out. This strain, which we will refer to as DM (for Double Mutant) from now on, was extensively characterized in our previous study[29] and shown to be devoid of 5mC and 6mA in CmCWGG and GmATC contexts respectively. Further, we enzymatically treated the DNA from DM with M.SssI and M.CviPI methylases to introduce 5mC in CpG and GpC contexts respectively. For *H.pylori*, we sequenced both the native (NAT) as well as whole-genome amplified (WGA) versions of the DNA. All the above samples were sequenced using both 4kHz and 5kHz sampling rates. We also generated bisulfite data for *E.coli* samples as ground truth. All bacterial data were sequenced to a target coverage of 200x. To ensure consistency across comparisons, we downsampled all human data to an average coverage of 50x, and all bacterial data to an average coverage of 100x (Fig S1A). In addition to the above, we used 5kHz data generated from two clinical isolates of *Pseudomonas aeruginosa* and *Staphylococcus aureus* to analyze specific contexts of 6mA and 4mC. The datasets used in this study, and the expected methylation contexts are outlined in Table S1.

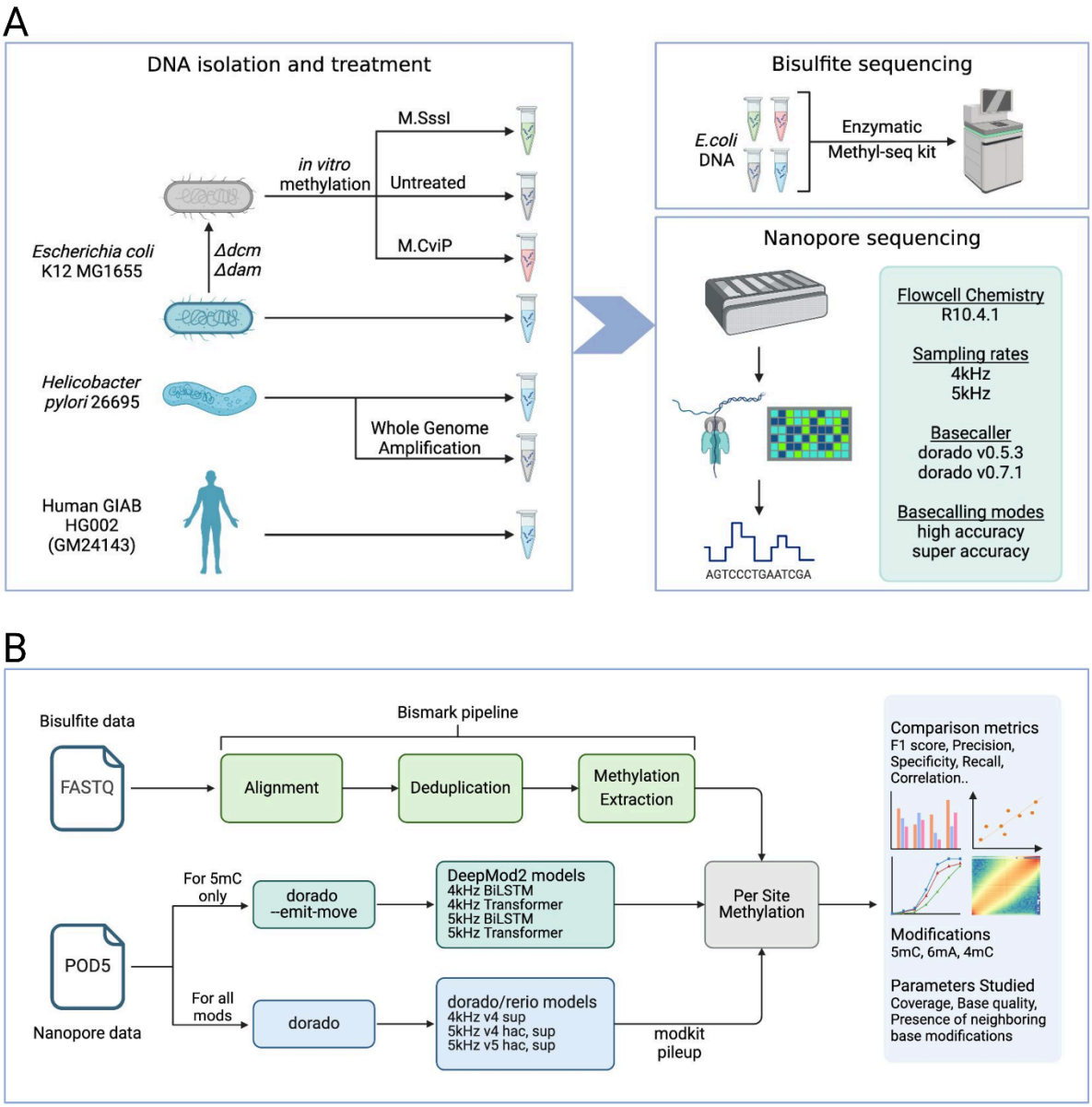


Figure 1: Benchmarking strategy. **A)** Whole genome sequencing workflow employed in this study. Wildtype/mutant/*in vitro* methylated/PCR amplified genomic DNA was sequenced on R10.4.1 flowcells at both 4kHz and 5kHz sampling rates. A ground truth for 5-methylcytosine was also generated for *E.coli* using EM-seq. **B)** Data analysis pipelines. Whole genome bisulfite data was processed following the Bismark pipeline. Raw POD5 files from nanopore sequencing were processed using Dorado with appropriate flags compatible with downstream processing. The DNA modifications studied, and the parameters used for evaluating the performance are indicated on the right. This figure was created using BioRender.com.

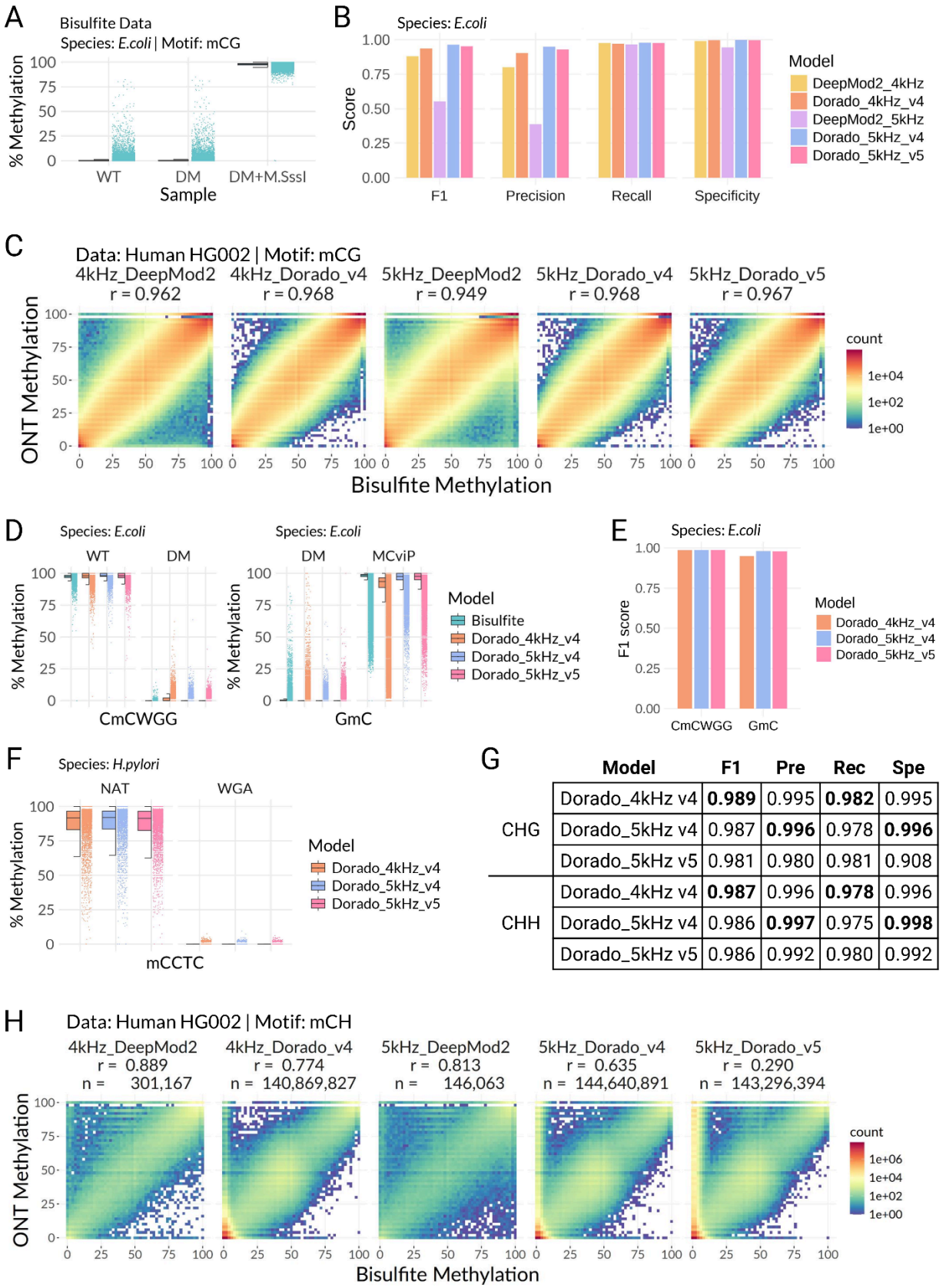
Most of the existing tools for methylation detection from ONT data work on the R9 chemistry, and have been benchmarked before[24,25]. As we focussed on the newer R10 chemistry, we limited our analysis to those tools which support methylation calling for this version. For 5mC, these are DeepMod2, Rockfish, and the models provided by ONT via Rerio and Dorado. For

ONT models, we chose the appropriate model for 4kHz data, as well as both v4 and the recently released (May 2024) v5 models for 5kHz data, in both high- (hac) and super-accuracy (sup) modes. For other modifications, currently no tool other than those offered by ONT exist. Hence, our analysis is limited to comparison of different versions of the ONT models for 6mA and 4mC . The average read quality of various basecalling models is depicted in figures S1B-S1D.

Comparing 5mC Methylation calls in CpG context

We first evaluated the efficacy of various tools and models in the most widely used setting: 5mC detection in CpG (mCG) context. Other than the models provided by ONT, currently CpG methylation calling for the R10 chemistry is supported by two tools: DeepMod2 and RockFish. Despite numerous attempts, we could not run RockFish on either our or publicly available datasets. Therefore, we excluded it from further analysis. DeepMod2 supports CpG calling for both 4kHz and 5kHz datasets. They also offer two kinds of models, one which uses the older bidirectional long short-term memory (BiLSTM) architecture and the other which uses new transformer models[23]. For both sampling rates, the BiLSTM models of DeepMod2 clearly outperformed the transformer models (Fig S2A). Hence, we used the BiLSTM models for all further comparisons. For Dorado, we used their all-context 5mC models, and subset the CpG locations for comparison with DeepMod2. For 4kHz, only the sup model is available for all-context 5mC calling whereas both hac and sup models are available for 5kHz sampling rate. Unsurprisingly, the sup models were superior (albeit slightly) to the corresponding hac models (Fig S2B), and therefore, unless explicitly mentioned, all comparisons were done with the sup models.

We used whole genome bisulfite sequencing (WGBS) data as the ground truth. WGBS indicated that the gDNA of *E.coli* DM strain was completely unmethylated at CpG sites, whereas the same gDNA treated with M.SssI showed near complete methylation (Fig 2A). On this data sequenced at 4kHz sampling rate, both DeepMod2 and Dorado v4 models performed extremely well in identifying 5mC (Fig 2B, S3A), with F1 scores of 0.88 and 0.93 respectively. Notably, Dorado models tailored for 5kHz data surpass their 4kHz counterparts in identifying mCG. The recall of all the five models was >0.96. However, the performance of the DeepMod2 5kHz model was worse compared to their 4kHz counterpart (F1 score 0.55 vs 0.88), owing to its lower precision (Fig 2B). A similar trend was observed with data of *H.pylori* in the sequence context GmCGC. Both 5kHz_v4 and 5kHz_v5 models of Dorado outperformed the 4kHz_v4 model in identification of 5mC, whereas the performance of DeepMod2 models was inferior (Fig S3B). In particular, all the models were extremely accurate in identifying unmethylated cytosines in this context, with the exception of the 5kHz model of DeepMod2.



G

	Model	F1	Pre	Rec	Spe
CHG	Dorado_4kHz v4	**0.989**	0.995	**0.982**	0.995
	Dorado_5kHz v4	0.987	**0.996**	0.978	**0.996**
	Dorado_5kHz v5	0.981	0.980	0.981	0.908
CHH	Dorado_4kHz v4	**0.987**	0.996	**0.978**	0.996
	Dorado_5kHz v4	0.986	**0.997**	0.975	**0.998**
	Dorado_5kHz v5	0.986	0.992	0.980	0.992

Figure 2: Performance of 5mC models. **A)** Jitter-box plot showing methylation percentage for *E.coli* bisulfite data. **B)** F1, precision, recall and specificity scores of all 5mC models in mCG context based on bisulfite data as ground truth. **C)** Heatmap depicting the correlation between methylation values called by each model in CG context and the expected ground truth (bisulfite data) for HG002 dataset. Locations depicted here (n = 14,839,407) are those commonly reported by all the tools as well as available in bisulfite data. The model name and Pearson correlation are indicated in the plot title. Color scale indicates the number of locations falling under each intersection bin. Bins with no values are blank. **D)** Performance of 5mC models in non-CG context using bisulfite data as ground truth for *E.coli* data, depicted as box-jitter plots. **E)** F1 score evaluation of 5mC models in non-CG context based on *E.coli* bisulfite data. **F)** Performance of Dorado models in non-CG (mCCTC) context evaluated on the native (NAT) and whole genome amplified (WGA) data of *H.pylori* **G)** Evaluation table showing F1 score, precision, recall and specificity for all Dorado models on the HG002 data in CHG and CHH contexts. **H)** Heatmap showing the correlation between methylation values called by each model in mCH (H = A/C/T) context and bisulfite data for HG002 dataset. The model name, Pearson correlation and total number of locations profiled are indicated in the plot title. Color scale indicates the number of locations falling under each intersection bin. Bins with no values are blank.

We further evaluated the performance of these models on human HG002 GIAB data, using the publicly available bisulfite data as ground truth. Overall, DeepMod2 models report more CpGs than the corresponding Dorado models, however, the Dorado models showed better correlation with bisulfite data (Fig S4A). To enable a better comparison, we have limited the analysis to a subset of 14.83M CpGs which were commonly profiled by all the 5 models and had at least a read coverage of 20. As can be seen in Figure 2C, all the models showed high correlation with bisulfite data, with Pearson r in the range of 0.962-0.968, with the exception of the DeepMod2_5kHz model which showed a slightly lower correlation of 0.949. Though the correlation values of DeepMod2 and Dorado models were similar, the output of Dorado was closer to the ground truth, as evidenced by empty areas in the top left and bottom right areas of the heatmap - areas which represent extreme disagreement (Fig 2C). When profiled for the exonic CpGs alone (n = 1,492,063), the correlation was even better: all models, including hac models of Dorado and the exception of DeepMod2_5kHz model, showed a Pearson r >0.98 (Fig S4B). Of note, the hac Dorado model of 5kHz performed on par with the sup model of 4kHz model. We then plotted the average CpG methylation around transcription start sites (TSS, n = 36,914). All the models show a characteristic dip at the TSS sites, closely following the trend of WGBS data (Fig S4C). Overall, we observed that identification of 5-methylcytosine in CpG context from Oxford Nanopore data is highly reliable.

Comparing 5mC Methylation calls in non CpG context

Though not as well studied, 5mC is also found in sequence contexts other than CpG. For example, in most bacterial systems, motifs known to generally harbor 5mC are not in CpG context[15]. In mammals, tissues such as brain harbor as much as 4-5% non CpG methylation, with several thousand sites shown to have close to 100% methylation[30]. Non CpG methylation has also been shown to play roles in development of cancer[31].

We evaluated the accuracy of various models to identify 5mC in contexts other than CpG. As DeepMod2 did not support 5mC calling in non CpG context, we limited our analysis to various “all-context” 5mC models of Dorado. *dcm* in WT *E.coli* methylates the second cytosine in the sequence motif CmCWGG (W = A/T). As can be seen in the left panel of Fig 2D, all the models

performed remarkably well in identifying methylated C in this context, on par with the bisulfite data. The 5kHz models were better than others in identifying unmethylated C in this context, as can be seen with the DM data (Fig 2D). We then tested the models on DM gDNA treated with M.CviP, which specifically methylates cytosines in GpC context. As expected bisulfite data shows close to 0% methylation in DM data, and ~100% methylation in DM+M.CviP data (Fig 2D, right). All models were remarkably well at identifying unmethylated C's in this context. However, the 5kHz models were better at identifying methylated C in GpC context compared to the 4kHz model (Fig 2D, right). Overall, the F1 scores for the all the models in both CmCWGG and GmC contexts was ~0.98, with the slight exception of 4kHz model in GmC context (F1 score of 0.95) owing to its lower recall (Fig 2E, Figs S5A-S5C). Surprisingly, the performance of these models was not as good in mCCTC context when tested on *H.pylori* data (Fig 2F). Further analysis revealed that this could be attributed to the base that was preceding the methylated C - All models performed substantially poorly in the TmCCTC context as compared to the other 3 bases (Fig S5D). Out of the 3, the v5 model of 5kHz showed the least bias in this context.

Next, we assessed model performance on the human HG002 dataset. Non-CpG methylation in mammals is usually analyzed as CHG (symmetrical) and CHH (hemi-methylated) contexts (H = A/T/C). Hence, we first chose to specifically group the non-CpG contexts into these two bins. As the number of cytosines unmethylated in these contexts is much higher than that of methylated cytosines, F1 score calculations on all cytosines will disproportionately penalize false negatives over false positives. Hence, to enable a fair comparison, we filtered locations to retain those cytosine-contexts that were reportedly 100% methylated in bisulfite data. This number was 4980 and 10823 in CHG and CHH contexts respectively. We then picked the same number of unmethylated cytosines (0% methylation according to bisulfite) randomly. Figure 2G depicts the precision, recall, sensitivity and F1 scores calculated using this subset of cytosines. All the 3 models performed remarkably well in these contexts. Out of the 3, the 4kHz_v4 model showed marginally better F1 and Recall values, whereas the 5kHz_v4 showed the best Precision and Specificity values compared to the others(Fig 2G). It was surprising that the latest v5 model did not outperform the older models in any of the 4 metrics. To expand the analysis beyond fully methylated or unmethylated cytosines, we selected all C's in non CpG context that showed >=1% methylation according to bisulfite (n = 145M). Figure 2H depicts the correlation of the output by various models with WGBS data. As expected for a tool that works in CpG context, DeepMod2 profiled very few C's in this context, but showed good correlation with bisulfite results for the few that it did profile (4kHz: r = 0.89, 5kHz: r = 0.81). In contrast, while Dorado models could profile most of the C's, their correlation was poorer. In fact, 5kHz_v5 had a r of 0.29, compared to 4kHz_v4 with an r of 0.77. This poor performance could be attributed to a lot of false positive calls, as evident by the high number of C's mapping to the left side of the heatmap (Fig 2H). Our results indicate that the performance of the all-context 5mC models was not as good in non-CpG contexts, particularly when identifying unmethylated cytosines.

Performance of 6mA models

In comparison to 5mC, tools that can identify 6mA from nanopore data have always been limited. This is particularly true for the new R10 chemistry data. As of writing this manuscript, there are no models to study 6mA from R10 chemistry other than those offered by Oxford

Nanopore Technologies. A possible reason for this is perhaps the lack of methods to generate ground truth data such as bisulfite data for 5mC.

ONT models for 6mA have been validated on synthetic oligo data, which may not reflect their performance on real datasets. Hence, we assessed various versions of ONT models on both *E.coli* and *H.pylori* data, the species for which general 6mA contexts have been studied and reported previously (Table S1). In GmATC context on *E.coli* data, the 5kHz models performed better than the 4kHz model, and the recent v5 model showed better performance than v4, particularly in identifying unmethylated adenines (Fig 3A). The models could also identify 6mA well in the context of AmACNNNNNNGTGC, which remains unaffected by the *dam* deletion. However, the performance of the v5 model seemed poorer than that of the v4 model on the 5kHz data in this context (Fig S6A).

We went on to characterize these models in various sequence contexts using native and WGA data of *H.pylori*. In general, the 5kHz_v4 model performed better than the 4kHz_v4 model. However, the most recent 5kHz_v5 model was worse at discriminating 6mA, particularly in identifying methylated adenines. Surprisingly, this was true even for GmATC context in *H.pylori*, indicating a possible species bias of the models (Fig 3B). In most cases, the 5kHz_v5 model retained its specificity, performing on par with the 5kHz_v4 model in identifying unmethylated adenines, but the recall was lower, similar to that of the 4kHz_v4 model (Fig 3D, Fig S6C). We further validated the v4-v5 differences on 5kHz data generated from two other species - clinical strains of *P.aeruginosa* and *S.aureus*. We observed a similar trend - the v5 model was not as good as the v4 model in identifying methylated adenines, although the difference is not equally prominent in all the tested contexts (Fig S7). Overall, our results indicated that except for GmATC context and specifically in *E.coli*, the v4 rather than the v5 models of 5kHz data performed the best in identification of 6mA in various sequence contexts.

Detection of 4-methylcytosine (4mC)

Though prevalent and mostly studied in bacteria, 4mC has been shown to also be present in few eukaryotic genomes[11]. Support for identification of 4mC from nanopore data is very recent, and models offered by ONT are currently available only for the 5kHz sampling rate. No other standalone tool at present supports 4mC calling. Hence, using data from *H.pylori*and a clinical strain of *P.aeruginosa,* we studied the accuracy of both the v4 and v5 4mC models offered by ONT. As with other models, both the versions were highly accurate in identifying unmethylated cytosine in the sequence contexts studied. In the T4mCCTC context for *H.pylori,* the median methylation detected was ~70% for both v4 and v5 models (Fig 3E, left). However, the same was >90% for GAC4mCCY and G4mCCNNNNNGTAG motifs from *P.aeruginosa*, indicating that the models have a species/context bias (Fig 3E, middle). Surprisingly, while no differences were observed between v4 and v5 models from *H.pylori* data, the v5 model was inferior in detecting 4mC for both the sequence contexts studied in *P.aeruginosa*.

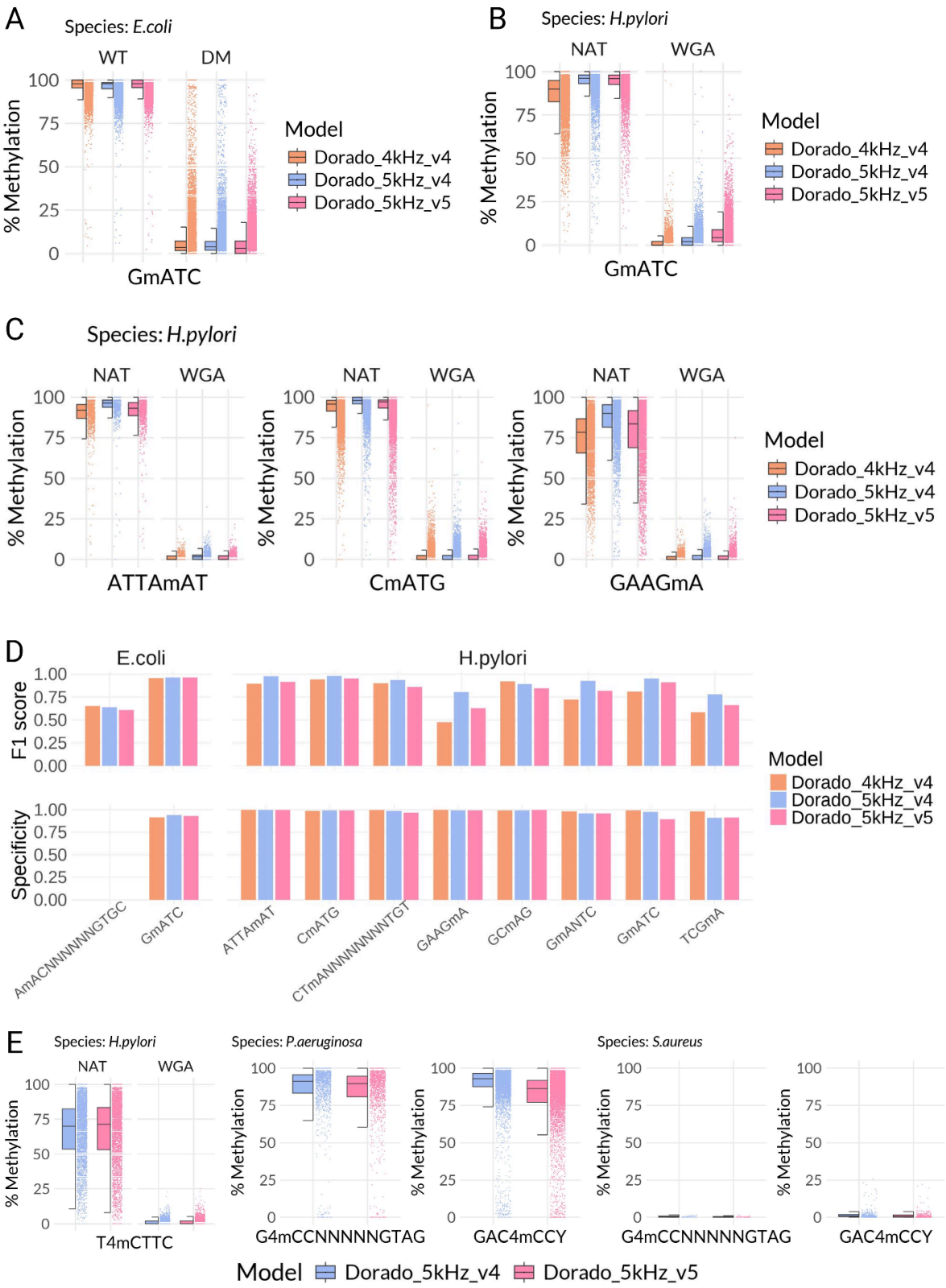


Figure 3: Performance of 6mA and 4mC models. **A)** Performance of 6mA models in GATC context for *E.coli* dataset. **B)** Performance of 6mA models in GATC context for *H.pylori* dataset. **C)** Performance

analysis of the 6mA models across various sequence contexts for the *H. pylori* datasets. **D)** F1 score (top panel) and Specificity (bottom panel) evaluation of 6mA models on diverse sequence contexts where methylation is expected in *E.coli* and *H.pylori* datasets. **E)** Performance of 4mC models in various contexts for *H.pylori* (left), *P.aeruginosa* (middle) and *S.aureus* (right) datasets.

Effect of sequencing coverage and quality

To check the effect of sequencing depth on the accuracy of methylation calls, we generated subsampled datasets ranging from an average of 5x to 70x coverage from both *E.coli* and *H.pylori* datasets. To account for possible subsampling biases, particularly in low coverage datasets, we performed 5 subsamples at each coverage value, and aggregated the result. In the previous sections, we already established the limitations and sequence specific biases of various models, particularly for identification of 6mA. To focus more on the effect of coverage rather than these limitations, we restricted our analysis to those locations which were called to be near completely methylated or unmethylated in the full dataset (100x), and used these locations for calculating the performance metrics (See Methods for details).

Figure 4A shows the effect of sequencing depth on the performance of various tools and models in CpG context. In line with the full dataset, DeepMod2 model for the 5kHz performed poorly across the coverages. Notably, Dorado models for both 4kHz and 5kHz achieved F1 scores >0.9 even at the lowest depth of 5x. Both v4 and v5 models of 5kHz performed better than the 4kHz_v4 model across all the coverages, and within the 5kHz models, the v4 version was marginally but consistently better than the v5 model. Interestingly, this trend was reversed in the case of CmCWGG context, where 5kHz_v5 performed better than 5kHz_v4 at all coverages (Fig 4B). However, the F1 scores approached >0.9 only at 10x coverage for this non-CpG context.

For 6mA, when profiled for GmATC context in *E.coli* data, all the 3 models performed comparably, with the latest 5kHz_v5 being slightly better at low coverages. All the models achieved an F1 score of ~0.95 at 20x depth. However, the same performance was not seen in GmATC context when tested on *H.pylori* data. Similar to what we observed with the full data, the performance of 5kHz_v4 was better than 5kHz_v5, particularly at low coverages (Fig 4C). At 20x coverage, only the 5kHz_v4 model approached an F1 score of 0.95. In fact, this was the case with all the contexts tested on *H.pylori* data (Fig 4D). Unlike for 5mC, we observe a clear jump in the performance metrics at 20x compared to 10x, after which they slowly plateau.

A major improvement with the increase in the sampling rate as well as the release of the new transformer-model architecture by ONT is the increase in the modal q-score (Fig S1B-D). Hence, we assessed if the confidence and accuracy of base modification detection improves as a function of the q-score. To this end, we filtered the reads using various minimum q-value thresholds (q8 to q22). For 4kHz data, we had to stop at q18, as the data loss beyond q18 was too high to be useful for comparative analyses. We tested the effect of minimum q-score in two contexts (mCG and CmCWGG) for 5mC, and one (GmATC) for 6mA. As can be seen in figure 4E, while the trend is slightly upward, the difference in the F1, precision, recall and specificity scores of the models tested was negligible. Therefore, our results indicate that the accuracy of base modification detection is not dependent on the q-scores of the underlying reads.

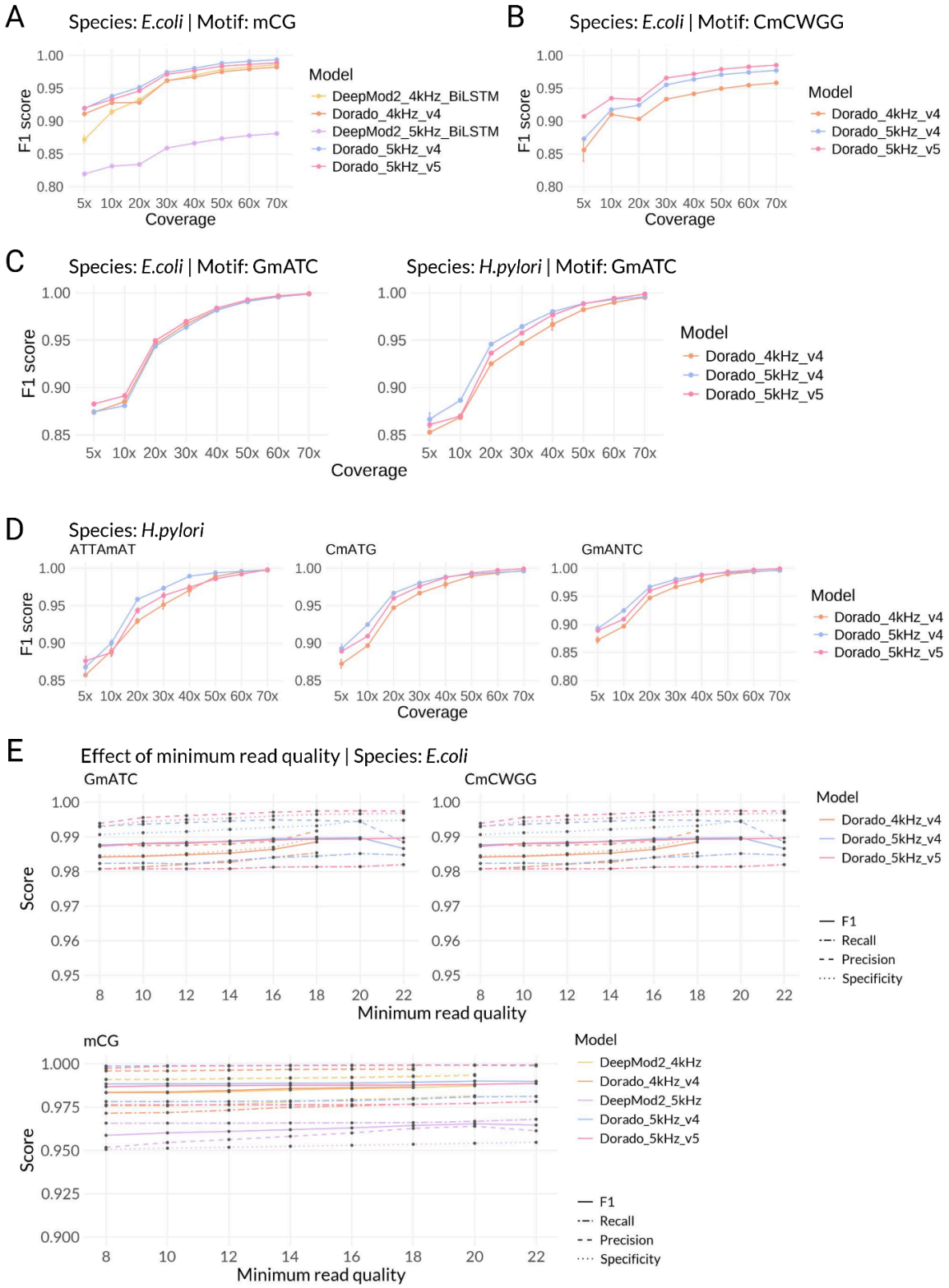


Figure 4: Effect of coverage and read-quality on methylation model performance. **A)** F1 score evaluation of all 5mC models in CG context for *E.coli* dataset subsampled to coverage values ranging from 5x to

70x. **B)** F1 score evaluation of Dorado 5mC models in non-CG (CCWGG) context for *E.coli* dataset subsampled to coverage values ranging from 5x to 70x. **C)** Comparison of F1 scores for dorado 6mA models in GATC context for *E.coli* and *H.pylori* datasets subsampled to coverage values ranging from 5x to 70x. **D)** F1 score evaluation of Dorado 6mA models in diverse sequence contexts for *H.pylori* datasets subsampled to different coverages. **E)** Performance of all methylation models in different sequence contexts (GATC for 6mA, CCWGG and CG for 5mC) for *E.coli* datasets subsampled using minimum read quality cutoff ranging from 8 to 22.

Effect of neighboring modifications

Approximately 8-10 nt of DNA contribute to the fluctuation of electrical signals at a time due to the longer sensing region of the current generation nanopores[17]. Hence, a modified base present in the vicinity of the target base can influence the accuracy of methylation calls. To assess this, we profiled the performance of various models in identifying mC and mA neighboring known methylated contexts. To begin with, we profiled the methylation status of the first cytosine in the motif CmCNGG using WT and DM data of *E.coli*. Bisulfite data confirms that the first C is not methylated in this context (Fig 5A, left). However, the 4kHz model of Dorado picked a lot of false positive methylation, only in the context CmCWGG and in WT data(Fig 5A, left, Fig S9A). We observed a significant improvement in the false positive calls with 5kHz models, particularly the v5 models.

We further sought to understand the effect of neighboring CpG methylation on false positive rates. Surprisingly, unlike the CmCWGG context, we noticed unmethylated C's preceding mCG being reported as mC by all the 3 models (Fig 5A, right). Moreover, the 4kHz model had the least bias among the 3 models. This trend was worse-off when there were two succeeding methylated CpGs (mCGmCG), where the 5kHz_v5 model reports close to 80% methylation for unmethylated C's (Fig S9B). To establish that the false methylation calls were due to the neighboring methylation rather than some sequence bias, we calculated the 'false positive rate' (FPR) of these models as a function of the distance to neighboring CpG methylation. FPR here was defined as the number of C's reported at a lenient cutoff of >50% methylation where none is expected, divided by the total number of C's. Both the 5kHz models, and specifically the v5 model, showed high FPRs when the distance to mCG was 3nt, which gradually decreased as the distance to mCG increased (Fig 5B, top). The 4kHz model on the other hand showed false positives only when the distance was 1nt (Fig 5B, top). The FPR of the v5 model was particularly bad in mCGmCG context, showing that this version was highly sensitive to neighboring CpG methylation (Fig 5B, bottom). Further analysis showed that there was a strong bias towards false methylation calls when the bases between the profiled C and mCG were adenines (Fig S9C).

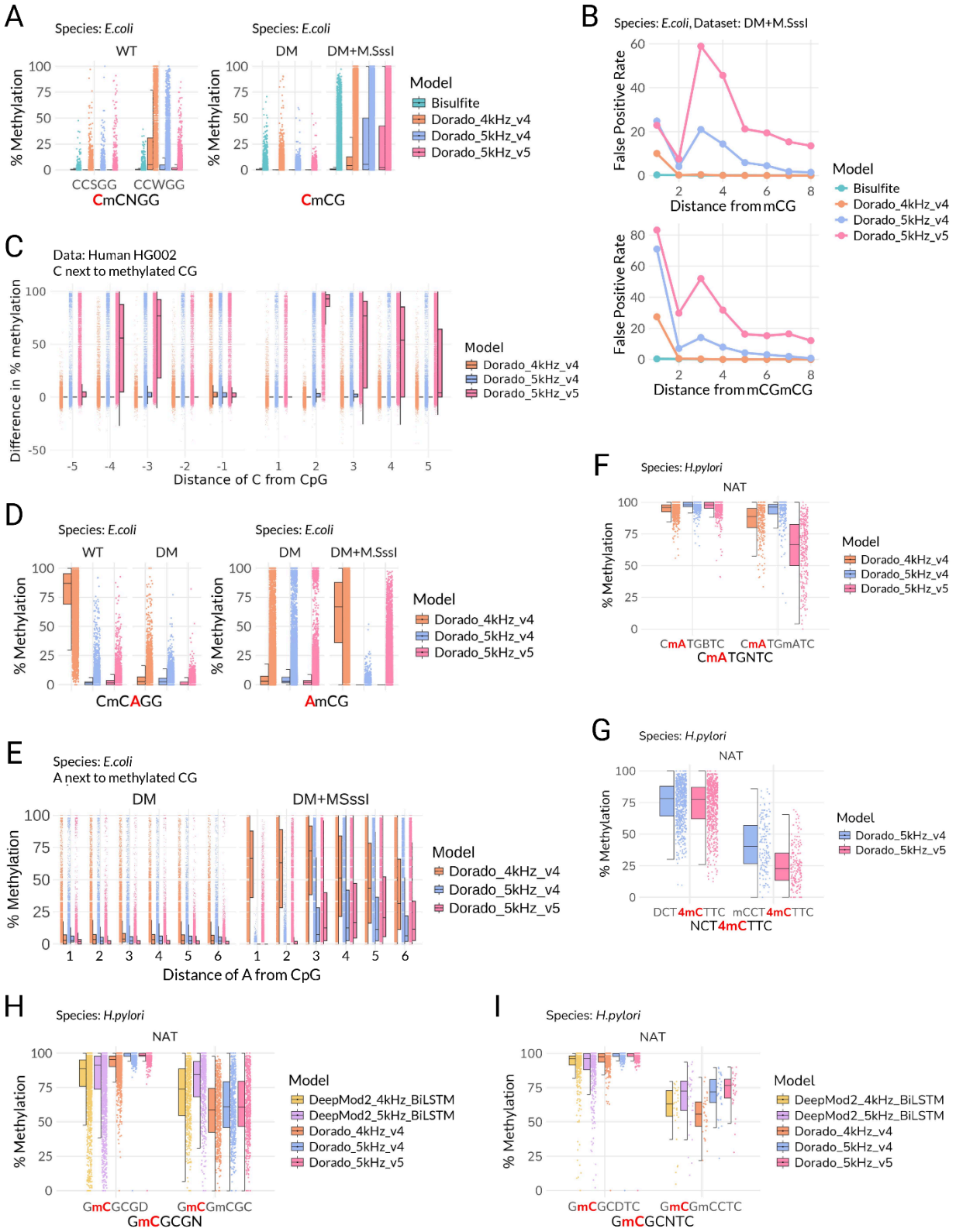

A
Species: E.coli
WT
Species: E.coli
DM
DM+M.SssI
% Methylation
CCSGG
CCWGG
CmCNGG
CmCG
Model
Bisulfite
Dorado_4kHz_v4
Dorado_5kHz_v4
Dorado_5kHz_v5
B
Species: E.coli, Dataset: DM+M.SssI
False Positive Rate
Distance from mCG
Distance from mCGmCG
C
Data: Human HG002
C next to methylated CG
Difference in % methylation
Distance of C from CpG
D
CmCAGG
AmCG
E
Species: E.coli
A next to methylated CG
DM+MSssI
Distance of A from CpG
F
Species: H.pylori
NAT
CmATGBTC
CmATGmATC
CmATGNTC
G
DCT4mCTTC
mCCT4mCTTC
NCT4mCTTC
H
DeepMod2_4kHz_BiLSTM
DeepMod2_5kHz_BiLSTM
GmCGCGD
GmCGmCGC
GmCGCGN
I
GmCGCDTC
GmCGmCCTC
GmCGCNTC

Figure 5: Effect of neighboring modifications on methylation model performance. **A)** Performance of Dorado 5mC models in identifying cytosine neighboring known methylated contexts. Bisulfite data confirms that the first C is not methylated in both contexts shown (CmCNGG,CmCG). **B)** Effect of neighboring CpG methylation on false positive rates. X-axis shows the distance of profiled C from the methylated C in mCG context (top panel) and mCGmCG context (bottom panel). **C)** Effect of neighboring CpG methylation in HG002 dataset. The Y-axis indicates the difference in percentage methylation predicted by the models compared to the bisulfite data. Negative values on the X-axis indicate that the profiled cytosine is upstream of the mCG, while positive values suggest it is downstream of the mCG. **D)** Performance of Dorado 6mA models in identifying unmethylated adenine neighboring known methylated contexts in *E.coli* dataset. Effect of methylated cytosine on succeeding adenine (left) and effect of methylated cytosine on preceding adenine (right). **E)** Performance of the Dorado 6mA models in detecting adenine neighboring CpG methylation over a distance of 1 to 6 bases in the *E. coli* dataset. The Y-axis displays the percentage methylation predicted by the models, while the X-axis indicates the distance in bases from the profiled adenine to the CpG in DM and M.SssI treated samples. **F)** Performance of the Dorado 6mA models in detecting mA neighboring adenine methylation in CmATGNTC context for *H.pylori* dataset. **G)** Performance of 4mC models in identifying methylation in absence (left) and presence (right) of neighboring methylated cytosine for *H.pylori* dataset. **H-I)** Performance of 5mC models in identifying methylation in absence (left) and presence (right) of neighboring methylated cytosine for *H.pylori* dataset in GmCGCGN (H) and GmCGCNTC (I) contexts.

To confirm this observation and to rule out possible species bias, we checked if this trend was true in the human HG002 data. For this, we selected all mCGs which were reported to be 100% methylated by WGBS. We then profiled neighboring C's in non-CpG context, and plotted the difference between the percent methylation reported by various ONT models versus that reported by WGBS. While the v4 models for both 4kHz and 5kHz perform relatively well, the latest 5kHz_v5 model significantly overestimated the methylation of C's neighboring methylated CpGs, confirming the false positive bias we observed with *E.coli* data (Fig 5C). We observed that even for human data, the bias is the strongest when the profiled C is 3nt away from the mC. We further asked if this bias was also seen for 6mA. Similar to C's, the false positive rate for unmethylated A's which are immediately next to methylated C's has improved a lot in the 5kHz_v5 model (Fig 5D). However, all the models picked up many false positives, starting from when the A is 3nt away from 5mC (Fig 5E). In this regard, the 5kHz_v4 showed the least false positives, followed by the 5kHz_v5 model. These results indicate that while the accuracy of modification calls immediately in the vicinity of other modified bases has improved a lot in the new models, their performance is worse for bases that are slightly further away.

We next asked if there was also an effect on the false negative rates due to the presence of neighboring modified bases. Using *H.pylori* data, we profiled the accuracy of methylation calls in the context CmATGNTC. We observed that the models showed false negatives when there was neighboring 6mA (CmATGmATC) but not in other contexts (Fig 5F). The false negative bias was the strongest in the 5kHz_v5 model, and least in the 5kHz_v4 model. Surprisingly, this bias was not seen when we profiled the opposite context: GmATCNTG (Fig S10A). Similarly, presence of neighboring 5mC affected the 4mC calls in the context 5mCCT4mCTTC (Fig 5G). We then asked if neighboring 5mC affected the accuracy of 5mC identification. All the models tested, including the DeepMod2 models, reported reduced methylation in the context GmCGCGN when there was a neighboring mCG present (Fig 5H). The false negative rates were the highest for the Dorado models. A similar trend was observed in the sequence context GmCGCNTC, where

the reported methylation drops to ~70% when there is a neighboring 5mC (Fig 5I). Overall, our results show that presence of neighboring base modifications affect both the false positive and false negative rates of various models, particularly when the modified base is slightly further away.

Discussion

Direct sequencing of the native, non-amplified DNA by long-read technologies such as ONT provides the advantage of simultaneous detection of base modifications present on the DNA. This advantage has been extensively leveraged so far to study DNA methylation using nanopore sequencing. Predictably, most of this work has revolved around CpG methylation, the best-studied methylation mark, and also the most prevalent in mammalian genomes. However, the real power of long-read sequencing is the ability to detect all DNA modifications in one go, as opposed to the traditional conversion based methods that capture only one modification at a time. This power is contingent on the availability of good models that can accurately identify various DNA modifications.

In our study, we show the current landscape of DNA base modification analysis using nanopore sequencing. The latest (v5) models offered by ONT indeed showcase a huge boost in q-scores compared to their previous models (Fig S1). However, their methylation detection capabilities may not always be concomitant with the q-score improvements. In fact, we show that the latest models offered by ONT perform similar to previous models for 5mC, but are worse in many cases for other marks, particularly in species such as *H.pylori*. This highlights the need to validate the performance on real data rather than synthetic controls that are typically used by ONT to check their models.

In many cases, particularly for non-CpG contexts and 6mA, we observed that the models performed well on *E.coli* data, but not on the corresponding *H.pylori* data. This indicates that the models might be exhibiting a species bias. This is further supported by the fact that the q-scores of *H.pylori* data are also on the lower side compared to that of Human and *E.coli* (Fig S1B-S1D). The underlying reason for this observation is yet to be determined, but a possible reason could be the GC% differences between *H.pylori* (~32%) and *E.coli* (~50%). Further experimentation is required to assess whether a basecalling model specifically trained for low GC genomes would perform better in such scenarios.

For CpG methylation, we have also assessed tools other than those offered by ONT. DeepMod2 has performed on par, though slightly worse, on the 4kHz data in comparison to the corresponding ONT model. However, their 5kHz model suffers from both false positives and false negatives. Of particular note is the performance of the hac Dorado models – they offer comparable accuracy to that of its sup counterpart, while running significantly faster. We propose that the accuracy offered by hac models is suitable for most CpG methylation studies, particularly for population-scale analyses.

As basecalling involves processing chunks of signals, it is expected that presence of neighboring modifications can influence the basecalling accuracy. Indeed, our hypothesis holds

true in our findings, particularly with the new v5 models. Interestingly we observe both improvement and deterioration in the new models; on the one hand, basecalls that are extremely close to methylated bases were highly accurate in the new models compared to their 4kHz counterparts. But the performance of the new models on bases that are three nucleotides or farther from methylated residues were worse than the older models. This was true for both false positives as well as false negatives. The new 5kHz_v5 models use a transformer architecture as compared to the traditional BiLSTM models. Similar deterioration in performance was also observed in the transformer models of DeepMod2. Given that transformer models are relatively new, these results suggest that there is scope for further improving model architecture and training, which can positively impact the accuracy of methylation calls.

In our work we also discuss the effect of sequencing metrics such as depth and read quality on the accuracy of base modification detection. For 5mC in CpG context, even as low as 5x coverage was adequate to achieve an F1 score of >0.9. However, for other marks such as 6mA, we required a minimum coverage of 20x to reach the F1 score of 0.9. Although not explicitly tested here, our findings suggest that the F1 scores for 15x coverage would be approaching 0.9. Furthermore, for 5mC we notice that the improvement in accuracy is gradual with increasing coverage. However, we observed a clear jump in F1 scores from 10x and 20x in the case of 6mA, indicating that calls for 6mA and perhaps 4mC will benefit more from having a higher data coverage. Interestingly the improvement of accuracy by filtering low quality reads was negligible, indicating that the Q-value of the underlying reads do not have a direct effect on methylation calling.

In this work we created extensive datasets from different bacterial species as well as the widely-studied – human HG002 sample to evaluate the performance of various methylation calling models. In addition to being useful to benchmark various tools, we believe that our datasets can also be leveraged by other researchers aiming to build their own models and network architectures for base modification detection from ONT data. Our study does have few limitations; firstly we did not benchmark 5-hydroxymethylcytosine (5hmC), another commonly studied modification. Also, while we could use WGBS data as ground truth for 5mC, no such data could be generated for other modifications such as 6mA and 4mC, and hence we had to rely on existing knowledge of various sequence contexts known to harbor these marks.

In summary, in this study we report the strengths and gaps of current state-of-the-art models available for studying DNA base modifications using nanopore sequencing. Our work will benefit researchers in the field seeking to understand how best to utilize the existing tools or models for their research to study DNA methylation. We also highlight the need for validating a tool or model on multiple datasets that can reliably showcase its performance across diverse scenarios.

Methods

Datasets

For *Escherichia coli* str. K12 substr. MG1655, we used the wildtype (WT) strain, as well as a *dam- dcm-* double mutant (DM) which lacks canonical 5mC and 6mA (in CmCWGG and GmATC contexts respectively). This DM strain was extensively characterized in our previous work[29]. For *Helicobacter pylori* str. 26695, we used the native (NAT) genomic DNA which harbors abundant adenine methylation (Table S1), 5mC, and 4mC. We generated its unmethylated counterpart by performing whole genome amplification (WGA). To generate more diverse sequence contexts specifically for 4mC and 6mA, we also sequenced a clinical isolate each of *P.aeruginosa* and *S.aureus* using the 5kHz sampling rate. For *E.coli*, we also performed whole genome bisulfite sequencing (WGBS) to establish the ground truth of 5mC. For the human data, we used the genomic DNA of HG002 GIAB cell line. The ground truth bisulfite data for HG002 was downloaded from the ONT public repository (see Data availability).

Genomic DNA extraction

Overnight cultures were washed in PBS and treated with a mixture of Lysozyme and SDS followed by Proteinase K treatment. Genomic DNA was purified using Cetyltrimethylammonium bromide (CTAB) followed by Phenol:Chlorofom:Isoamyl alcohol. The gDNA was then washed with Chloroform:Isoamyl amyl alcohol and precipitated using Isopropanol. The gDNA was spooled using a glass rod, washed with ethanol and air dried. The spooled gDNA was dissolved in EB (Qiagen, Germany) and quantified using Qubit (ThermoFisher Scientific, US).

in vitro DNA Methylation

For *in vitro* DNA methylation, the double mutant *E.coli* was used[29]. The gDNA was treated with M.SssI or M.CviPI (NEB, USA) for CpG or GpC methylation respectively. The gDNA was treated with the enzyme in a mix containing S-Adenosyl methionine (SAM) and its corresponding buffer, for 4 hours at 37°C. The samples were purified using Ampure XP beads (Beckman Coulter, USA) and quantified using Qubit (ThermoFisher Scientific, US).

Whole Genome Amplification

Genomic DNA from *H.pylori* was used as a template for whole genome amplification. To amplify the genetic material, the PCR Barcoding Expansion Kit (Nanopore, UK) was used according to the manufacturer's protocol. 100 ng of genomic DNA was A-tailed using NEBNext Ultra II End Repair / dA-tailing Module (NEB, US) and purified using Ampure XP beads (Beckman Coulter, US). Barcode adapters were ligated to the sample and purified. The samples were PCR amplified for 15 cycles using compatible PCR Barcodes from the kit and LongAmp Taq 2X master mix (NEB, US). The amplified product was purified, and subsequently quantified using Qubit (ThermoFisher Scientific, US).

Nanopore library preparation

Nanopore libraries were prepared using Native Barcoding Kit V14 (Oxford Nanopore, UK) according to the manufacturer's protocol. In brief, the gDNA or WGA DNA was treated with

NEBNext Ultra II End Repair/dA-Tailing Module and NEBNext FFPE DNA Repair Mix (NEB, US). The samples were purified using Ampure XP beads (Beckman Coulter, US) and barcodes were ligated to the DNA fragments. Samples were purified once more and quantified using Qubit (ThermoFisher Scientific, US). Equimolar amounts of the barcoded samples were pooled and ligated to the native adapter. Post purification, the library was quantified and 50 fmol was loaded onto an R10.4.1 MinION flow cell or an R10.4.1 PromethION flow cell and sequenced.

Bisulfite library prep

DNA was processed using the NEBNext Enzymatic Methyl-seq Kit (NEB, USA) as per the manufacturer's protocol. The DNA was fragmented using Covaris Focussed Sonicator to an average fragment size of 300 bp as measured by a TapeStation (Agilent, US). The fragmented DNA was treated with an end prep enzyme mix followed by adapter ligation. DNA was oxidized using a TET2 containing mix and denatured with Formamide. The denatured DNA was treated with an APOBEC mix to deaminate the Cytosines. Indexes were incorporated into the DNA fragments using PCR and purified afterwards. Post quantification, the samples were pooled at an equimolar ratio. The pooled library was run on an Illumina Novaseq 6000 at PE150.

Bisulfite Data Processing

The quality of Illumina raw reads was checked using FastQC. Adapters were trimmed and low quality reads were filtered using Cutadapt with default parameters[32]. The trimmed data was then mapped to the *E.coli* reference genome (*E.coli* str. K12 substr. MG1655) using Bismark[33]. The mapped data was deduplicated and cytosine methylation was extracted for all the cytosines using 'deduplicate_bismark' and 'bismark_methylation_extractor' respectively. Bisulfite data for the human HG002 cell line was downloaded from the ONT open data repository hosted on AWS (see Data Availability).

Basecalling and quality control

We used the Dorado basecaller[19] (v0.5.3) for basecalling all nanopore reads in POD5 file format using the v4 models. In May 2024, ONT released their new v5 models compatible with their 5kHz sampling rate datasets, which claim to have improved both the sequencing and the methylation calling accuracy. Hence, we also basecalled all our 5kHz data with the new v5 models using Dorado v0.7.1. For DeepMod2, we basecalled the POD5 datasets using Dorado with the '–emit-move' flag to generate BAM files that were compatible with downstream processing using DeepMod2.

Methylation Calling

For 4kHz datasets, we used the appropriate 4kHz model, along with the specific 5mC models for modified basecalling from the Rerio model repository (see Data Availability). 5kHz datasets were basecalled for modifications using 5kHz-specific v4 and v5 5mC, 6mA, and 4mC Dorado models. Moreover, since the 5mC 4kHz model is only available in super accuracy (sup), we basecalled 4kHz with the sup model while 5kHz datasets were basecalled with sup as well as the high accuracy (hac) models. Unaligned BAM files produced by Dorado were then converted to FASTQ files using 'samtools fastq'[34] with the '-T "*" ' argument to preserve all the read tags

included in the BAM file; these reads were then aligned to the appropriate reference genomes using Minimap2[35].

Quality checks were performed on the aligned-bam files using NanoStat[36] and the coverage of the dataset was calculated using mosdepth[37]. In order to eliminate possible confounding biases introduced by variations in coverage, all human data was down-sampled to an average depth of 50x, and all bacterial datasets to an average depth of 100x using Rasusa[38].

For DeepMod2, we used both BiLSTM and transformer models for 4kHz as well as 5kHz data. 'deepmod2 detect' was run on BAM files with move tables to generate per-site CpG methylation values. 'deepmod2 merge' was used to merge the per-read calls and obtain per-site calls. For Dorado, we used both hac and sup models wherever they were available. Aligned Modbam files generated by Dorado models were processed using Modkit[39]. 'modkit pileup' was used to derive per-site methylation values in BED format. In addition to the methylation percentage, Modkit also provides information about the read coverage as well as the number of supporting reads. To account for coverage biases, we removed all sites that were covered <20x (except when studying the effect of coverage). Comparison of per-site values with bisulfite data was done using custom Python scripts. The data was summarized and visualized using dplyr and ggplot2 in the R statistical environment.

To study the effect of neighboring modifications in humans, cytosines flanking five positions upstream and downstream of fully methylated (100%) and unmethylated (0%) mCGs, as indicated by WGBS, were profiled. For the E.coli dataset, adenines flanking up to six positions downstream of methylated cytosines were profiled. It was ensured that no unintended mCGs were included between the mCG and the profiled base. The difference between modkit and WGBS methylation percentages was plotted as a jitter-box plot.

Assessing CpG methylation calls in exons and transcription start sites

We downloaded the 45th release (GRCh38.p14) of the comprehensive gene annotation (on the reference chromosome only) file from GENCODE in GTF format (see Data availability). The coordinates of all exons were retrieved from this file, and stored as a BED file using an awk one-liner. All cytosines that were present in exons were fetched from the full genome files using bedtools[40]. Similarly, awk was used to fetch all the transcription start sites based on gene orientation and extended by 5 kb on either side. The methylation profiles of the entire 10 kb region for all the TSS (n = 36,914) were averaged and visualized as a line plot using R.

Calculating F1 scores

F1 score calculation was performed differently for different datasets, based on the species and the modification of interest. For 5mC, bisulfite data was available for both human as well as the *E.coli* datasets where Modkit and DeepMod2 outputs were intersected with the bisulfite data. An equal number of sites with 100% and 0% bisulfite methylation were filtered as positive and negative sets, respectively. The number of reads classified as methylated and unmethylated by the tools were used to count true positive (TP), true negative (TN), false positive (FP) and false

negative (FN) values. These values were used to calculate the F1 score, precision, recall, and specificity, employing the following formulae.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Studying the effect of coverage

The aligned *E.coli* BAM files for 4kHz and 5kHz data were subsampled to 5x, 10x, 20x, 30x, 40x, 50x, 60x and 70x coverage using the tool, Rasusa. The subsampled BAM files were used to generate new per-site methylation summary files using Modkit or DeepMod2. The random subsampling process was repeated five times for each dataset to account for possible subsampling artefacts.

For 5mC, the per-site files were intersected with the reference bisulfite data. Sites with more than 90% bisulfite methylation were considered to be truly methylated, whereas those with less than 10% methylation were classified as truly unmethylated. Sites not agreeing to the bisulfite calls were considered false positives and false negatives. These values were used to calculate F1, precision, recall, and specificity scores. In the absence of definitive ground truth data for 6mA, we used motif-specific loci from the original (100x) BAM file as a reference. Sites with over 85% methylation were classified as truly methylated, while those with less than 15% methylation were deemed truly unmethylated. The subsequent steps for calculating F1 scores, precision, recall, and specificity followed a similar process to that of 5mC analysis.

Evaluating the effect of read-quality

We limited the analysis to *E.coli* data as the overall quality of *H.pylori* data was low to begin with. The aligned BAM files for both 4kHz and 5kHz data were subsampled on the basis of a minimum read-quality threshold using Nanoq[41]. The minimum read-quality cut-offs were set to 8, 10, 12, 14, 16, 18, 20 and 22. The subsampled BAM files were reprocessed using Modkit or DeepMod2 to derive new per-site methylation values. For 5mC, bisulfite data was used to determine mCG as well as CmCWGG sites that were either fully methylated or unmethylated. Reads agreeing with the bisulfite calls were considered true positives and true negatives, whereas those not agreeing were assigned to false negatives and false positives. For 6mA in GmATC context, a 90% and 10% methylation cutoff was used to determine true positives and

true negatives respectively. These values were used to calculate F1, precision, recall, and specificity scores as mentioned in the previous sections.

Acknowledgements

The authors acknowledge Dr Santosh Kumar and Apoorva Etikala for the genomic DNA of *Helicobacter pylori* str. 26695. Valli Undamatla and Tulasi Nagabandi from the NGS facility of CCMB are acknowledged for help in Bisulfite sequencing. The authors thank Dr Khader Valli Rupanagudi and Dr Reddy P Kommaddi from Centre for Brain Research, Bengaluru, for providing the HG002 GIAB cell line.

Funding

Work in this study was financially supported by the Rockefeller Foundation grant 2021 HTH 018 and the grant BT/PR40264/BTIS/137/44/2022 by the Department of Biotechnology, Government of India.

Data availability

Reference aligned Modbam files generated in this study are deposited to the NCBI SRA under the BioProject number PRJNA1172918.

Other data sources:
hg38 gene annotations:
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_45/gencode.v45.annotation.gtf.gz
Rerio repository: https://github.com/nanoporetech/rerio
HG002 WGBS data: s3://ont-open-data/gm24385_mod_2021.09/bisulphite/bismark/

Conflict of interest statement

DTS has participated as a speaker in events organized by Oxford Nanopore Technologies. Other authors declare no competing interests.

References

1. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
2. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).
3. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
4. Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation

variation. *Nature* **523**, 212–216 (2015).

5. Zhou, W. *et al.* DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018).
6. Hofer, A., Liu, Z. J. & Balasubramanian, S. Detection, Structure and Function of Modified DNA Bases. *J. Am. Chem. Soc.* **141**, 6420–6429 (2019).
7. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: In the right place at the right time. *Science* **361**, 1336–1340 (2018).
8. Au, K. G., Welsh, K. & Modrich, P. Initiation of methyl-directed mismatch repair. *J. Biol. Chem.* **267**, 12142–12148 (1992).
9. Arber, W. & Dussoix, D. Host specificity of DNA produced by *Escherichia coli*: I. Host controlled modification of bacteriophage λ. *J. Mol. Biol.* **5**, 18–36 (1962).
10. Boulias, K. & Greer, E. L. Means, mechanisms and consequences of adenine methylation in DNA. *Nat. Rev. Genet.* **23**, 411–428 (2022).
11. Rodriguez, F., Yushenova, I. A., DiCorpo, D. & Arkhipova, I. R. Bacterial N4-methylcytosine as an epigenetic mark in eukaryotic DNA. *Nat. Commun.* **13**, 1072 (2022).
12. Krueger, F., Kreck, B., Franke, A. & Andrews, S. R. DNA methylome analysis using short bisulfite sequencing data. *Nat. Methods* **9**, 145–151 (2012).
13. Ji, L. *et al.* Methylated DNA is over-represented in whole-genome bisulfite sequencing data. *Front. Genet.* **5**, (2014).
14. Grunau, C., Clark, S. J. & Rosenthal, A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res.* **29**, e65 (2001).
15. Beaulaurier, J., Schadt, E. E. & Fang, G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nat. Rev. Genet.* **20**, 157–172 (2019).
16. Ip, C. L. C. *et al.* MinION Analysis and Reference Consortium: Phase 1 data release and analysis. Preprint at https://doi.org/10.12688/f1000research.7201.1 (2015).
17. Van der Verren, S. E. *et al.* A dual-constriction biological nanopore resolves homonucleotide

sequences with high fidelity. *Nat. Biotechnol.* **38**, 1415–1420 (2020).

18. Oxford Nanopore [@nanopore]. Improvements in latest MinKNOW update: •5kHz sampling rate replaces 4kHz sampling rate - for improved accuracy •POD5 is now the default raw data file type •Addition of 5mC/5hmC Remora Model •Automatically stop your runs with Run Until Learn more: https://bit.ly/3MFtbHf https://t.co/2Pizs3LP1z. *Twitter* https://x.com/nanopore/status/1659469919971360769 (2023).
19. nanoporetech/dorado. Oxford Nanopore Technologies (2024).
20. Transforming Basecalling in Genomic Sequencing. *Oxford Nanopore Technologies* https://nanoporetech.com/blog/transforming-basecalling-in-genomic-sequencing (2024).
21. *London Calling 2024: Update from Clive Brown*. (2024).
22. Stanojević, D., Li, Z., Bakić, S., Foo, R. & Šikić, M. Rockfish: A transformer-based model for accurate 5-methylcytosine prediction from nanopore sequencing. *Nat. Commun.* **15**, 5580 (2024).
23. Ahsan, M. U., Gouru, A., Chan, J., Zhou, W. & Wang, K. A signal processing and deep learning framework for methylation detection using Oxford Nanopore sequencing. *Nat. Commun.* **15**, 1448 (2024).
24. Yuen, Z. W.-S. *et al.* Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nat. Commun.* **12**, 3438 (2021).
25. Liu, Y. *et al.* DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biol.* **22**, 295 (2021).
26. Genner, R. *et al.* Assessing methylation detection for primary human tissue using Nanopore sequencing. 2024.02.29.581569 Preprint at https://doi.org/10.1101/2024.02.29.581569 (2024).
27. Sigurpalsdottir, B. D. *et al.* A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. *Genome Biol.* **25**, 69 (2024).
28. Tourancheau, A., Mead, E. A., Zhang, X.-S. & Fang, G. Discovering multiple types of DNA

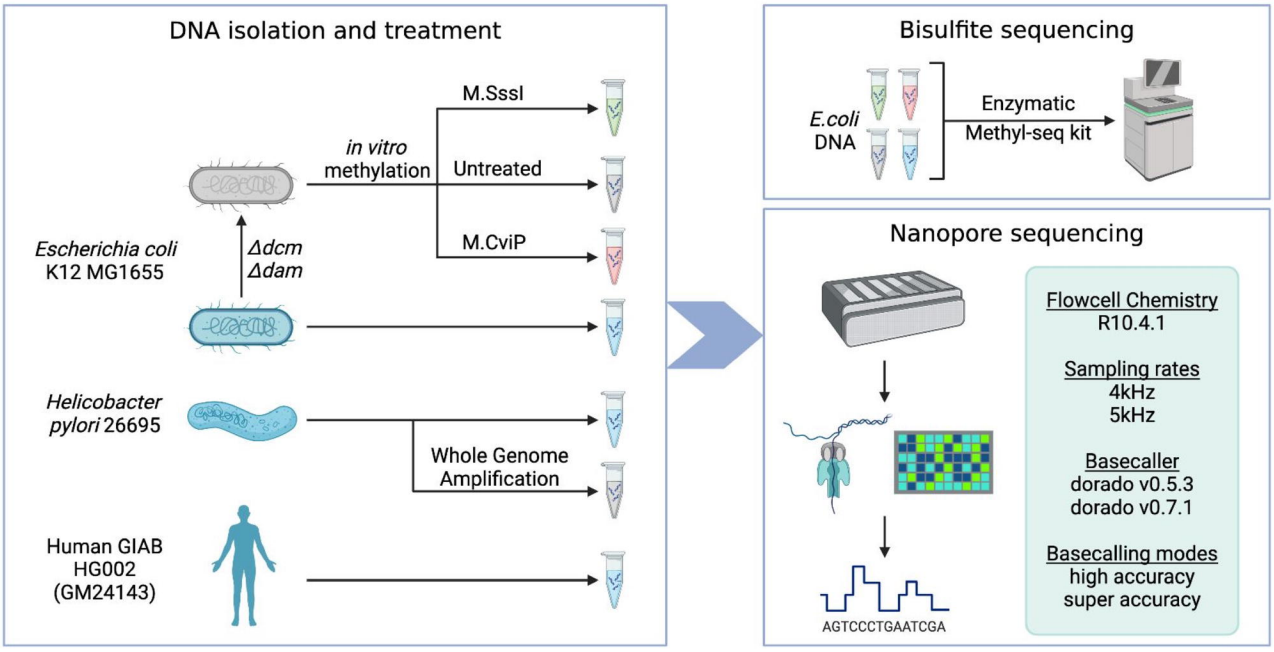
methylation from bacteria and microbiome using nanopore sequencing. *Nat. Methods* **18**, 491–498 (2021).

29. Kulkarni, O. *et al.* NEMO: Improved and accurate models for identification of 6mA using Nanopore sequencing. 2024.03.12.584205 Preprint at https://doi.org/10.1101/2024.03.12.584205 (2024).
30. Jang, H. S., Shin, W. J., Lee, J. E. & Do, J. T. CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function. *Genes* **8**, 148 (2017).
31. Ramasamy, D., Deva Magendhra Rao, A. K., Rajkumar, T. & Mani, S. Non-CpG methylation—a key epigenetic modification in cancer. *Brief. Funct. Genomics* **20**, 304–311 (2021).
32. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
33. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
34. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
35. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
36. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
37. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
38. Hall, M. B. Rasusa: Randomly subsample sequencing reads to a specified coverage. *J. Open Source Softw.* **7**, 3941 (2022).
39. nanoporetech/modkit. Oxford Nanopore Technologies (2024).

40. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
41. Steinig, E. & Coin, L. Nanoq: ultra-fast quality control for nanopore reads. *J. Open Source Softw.* **7**, 2991 (2022).

Figure 1

A


DNA isolation and treatment
M.SssI
in vitro methylation
Untreated
M.CviP
Escherichia coli K12 MG1655
Δdcm Δdam
Helicobacter pylori 26695
Whole Genome Amplification
Human GIAB HG002 (GM24143)
Bisulfite sequencing
E.coli DNA
Enzymatic Methyl-seq kit
Nanopore sequencing
AGTCCCTGAATCGA
Flowcell Chemistry
R10.4.1
Sampling rates
4kHz
5kHz
Basecaller
dorado v0.5.3
dorado v0.7.1
Basecalling modes
high accuracy
super accuracy

B

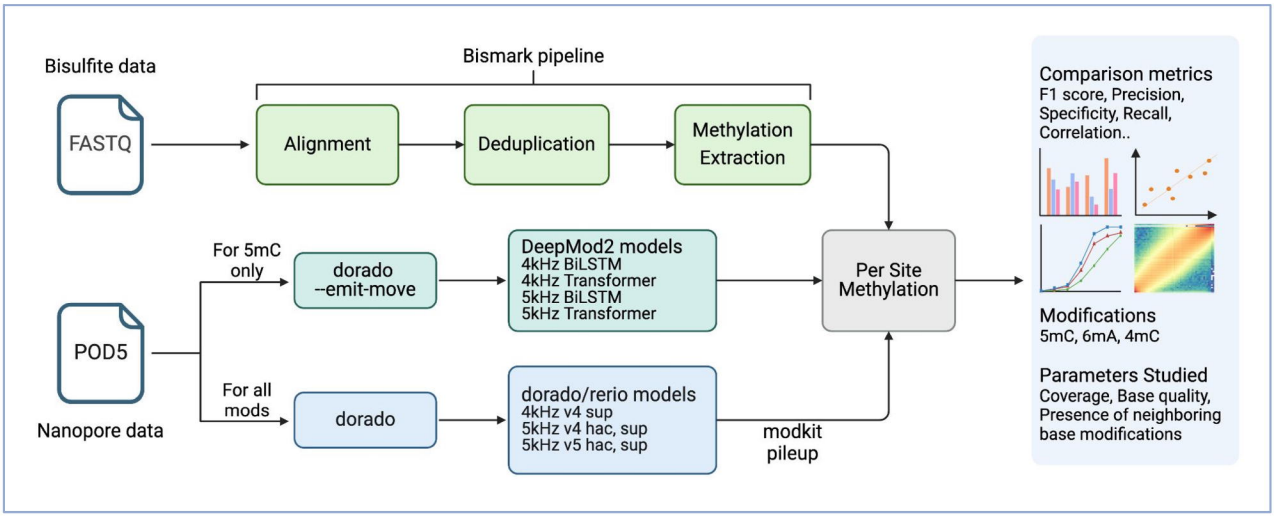
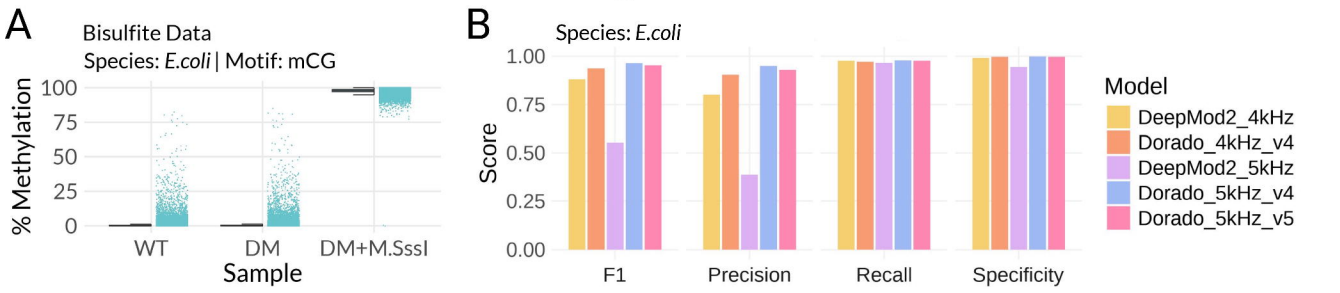
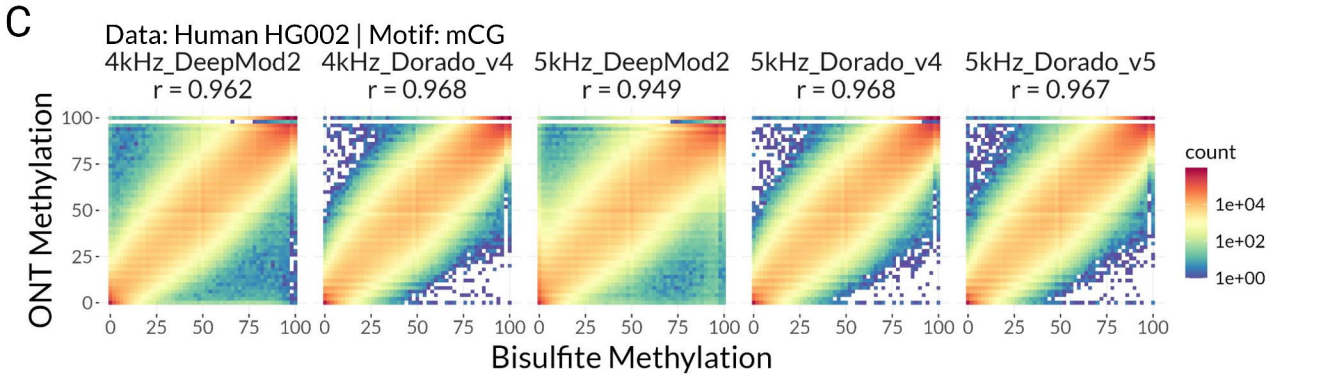
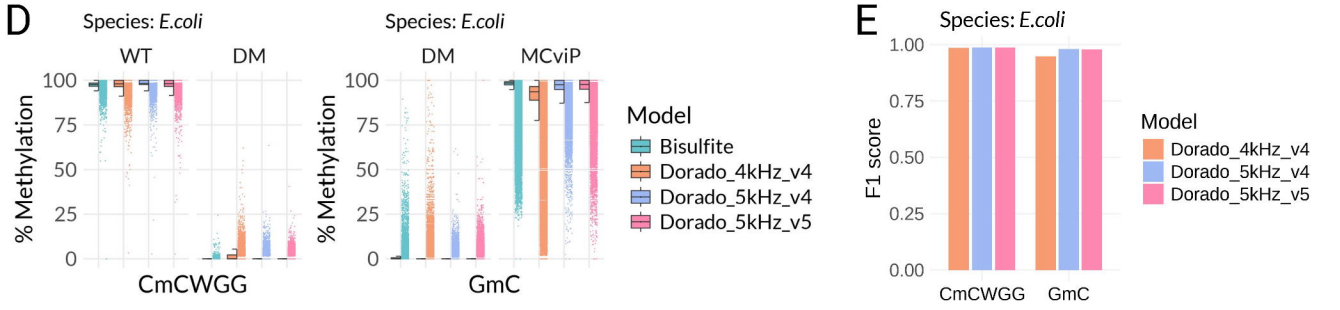
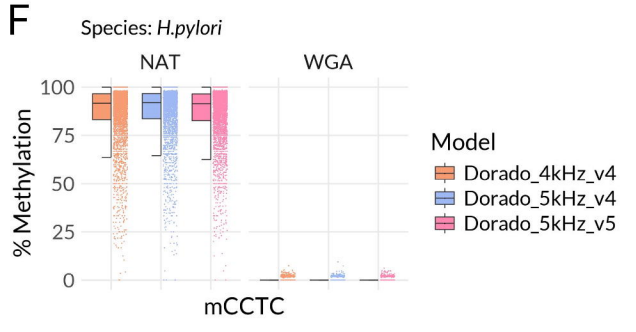

Bisulfite data
FASTQ
Bismark pipeline
Alignment
Deduplication
Methylation Extraction
POD5
Nanopore data
For 5mC only
dorado --emit-move
DeepMod2 models
4kHz BiLSTM
4kHz Transformer
5kHz BiLSTM
5kHz Transformer
For all mods
dorado
dorado/rerio models
4kHz v4 sup
5kHz v4 hac, sup
5kHz v5 hac, sup
modkit pileup
Per Site Methylation
Comparison metrics
F1 score, Precision, Specificity, Recall, Correlation..
Modifications
5mC, 6mA, 4mC
Parameters Studied
Coverage, Base quality, Presence of neighboring base modifications

Figure 2









G

	Model	F1	Pre	Rec	Spe
CHG	Dorado_4kHz v4	**0.989**	0.995	**0.982**	0.995
	Dorado_5kHz v4	0.987	**0.996**	0.978	**0.996**
	Dorado_5kHz v5	0.981	0.980	0.981	0.908
CHH	Dorado_4kHz v4	**0.987**	0.996	**0.978**	0.996
	Dorado_5kHz v4	0.986	**0.997**	0.975	**0.998**
	Dorado_5kHz v5	0.986	0.992	0.980	0.992

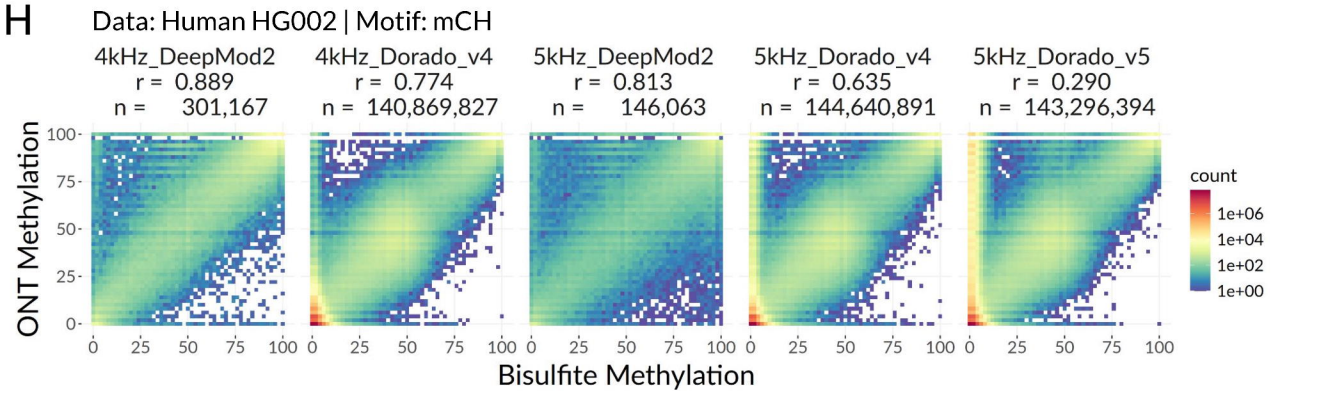


Figure 3

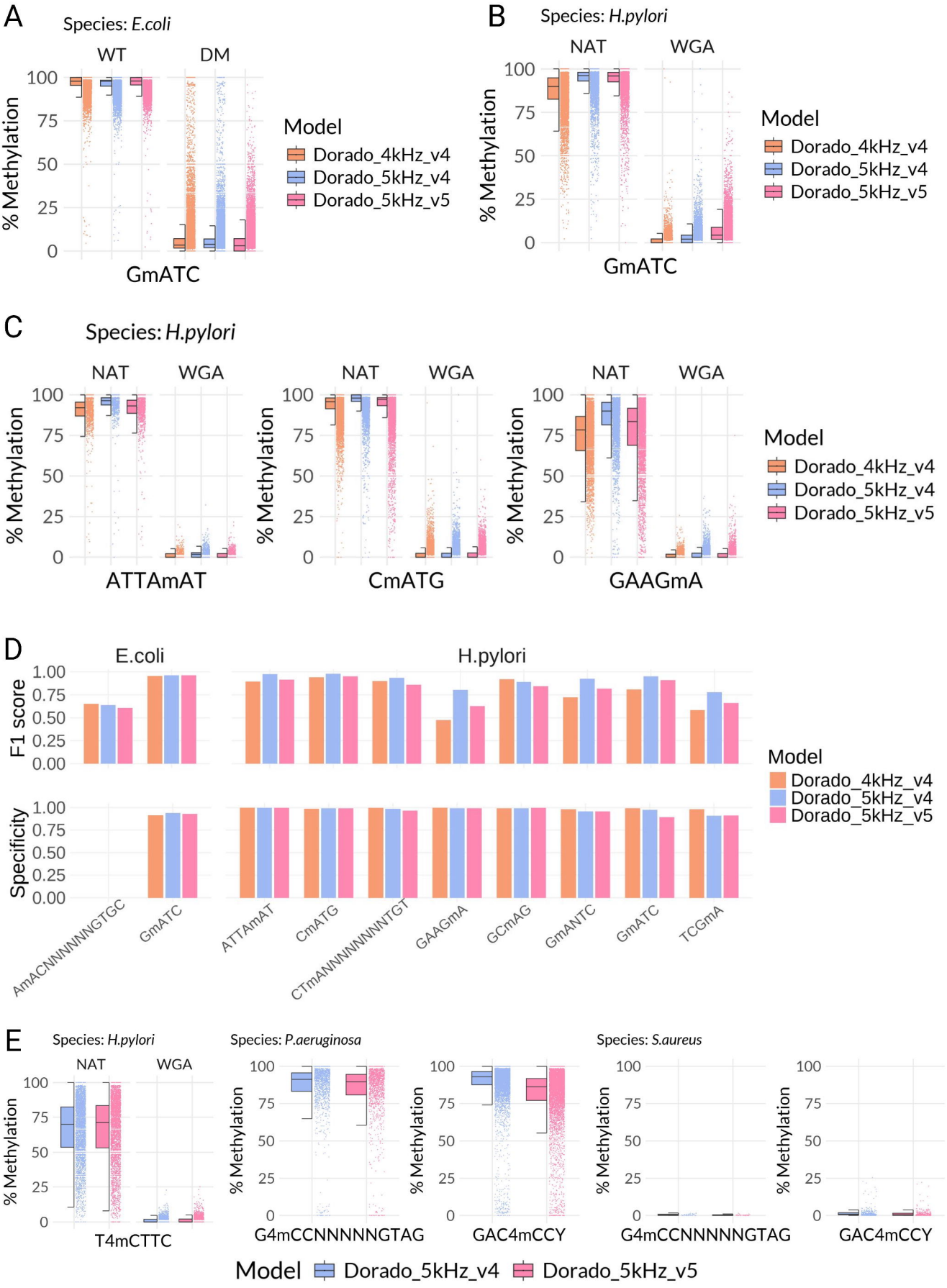

A
Species: E.coli
WT
DM
% Methylation
GmATC
Model
Dorado_4kHz_v4
Dorado_5kHz_v4
Dorado_5kHz_v5
B
Species: H.pylori
NAT
WGA
% Methylation
GmATC
Model
Dorado_4kHz_v4
Dorado_5kHz_v4
Dorado_5kHz_v5
C
Species: H.pylori
NAT
WGA
% Methylation
ATTAmAT
CmATG
GAAGmA
Model
Dorado_4kHz_v4
Dorado_5kHz_v4
Dorado_5kHz_v5
D
E.coli
H.pylori
F1 score
Specificity
AmACNNNNNGTGC
GmATC
ATTAmAT
CmATG
CTmANNNNNNNTGT
GAAGmA
GCmAG
GmANTC
GmATC
TCGmA
Model
Dorado_4kHz_v4
Dorado_5kHz_v4
Dorado_5kHz_v5
E
Species: H.pylori
NAT
WGA
% Methylation
T4mCTTC
Species: P.aeruginosa
% Methylation
G4mCCNNNNNGTAG
GAC4mCCY
Species: S.aureus
% Methylation
G4mCCNNNNNGTAG
GAC4mCCY
Model
Dorado_5kHz_v4
Dorado_5kHz_v5

Figure 4

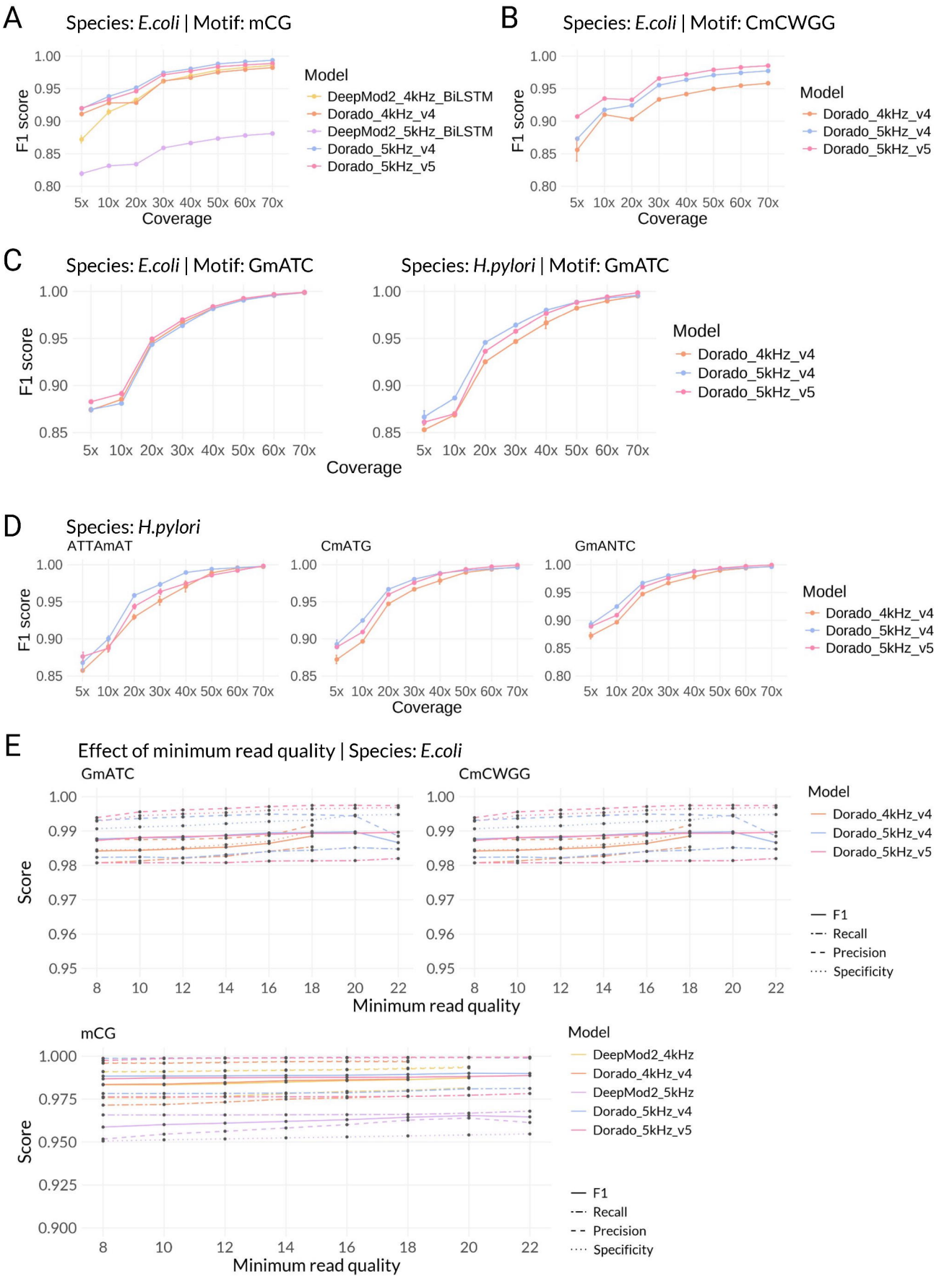
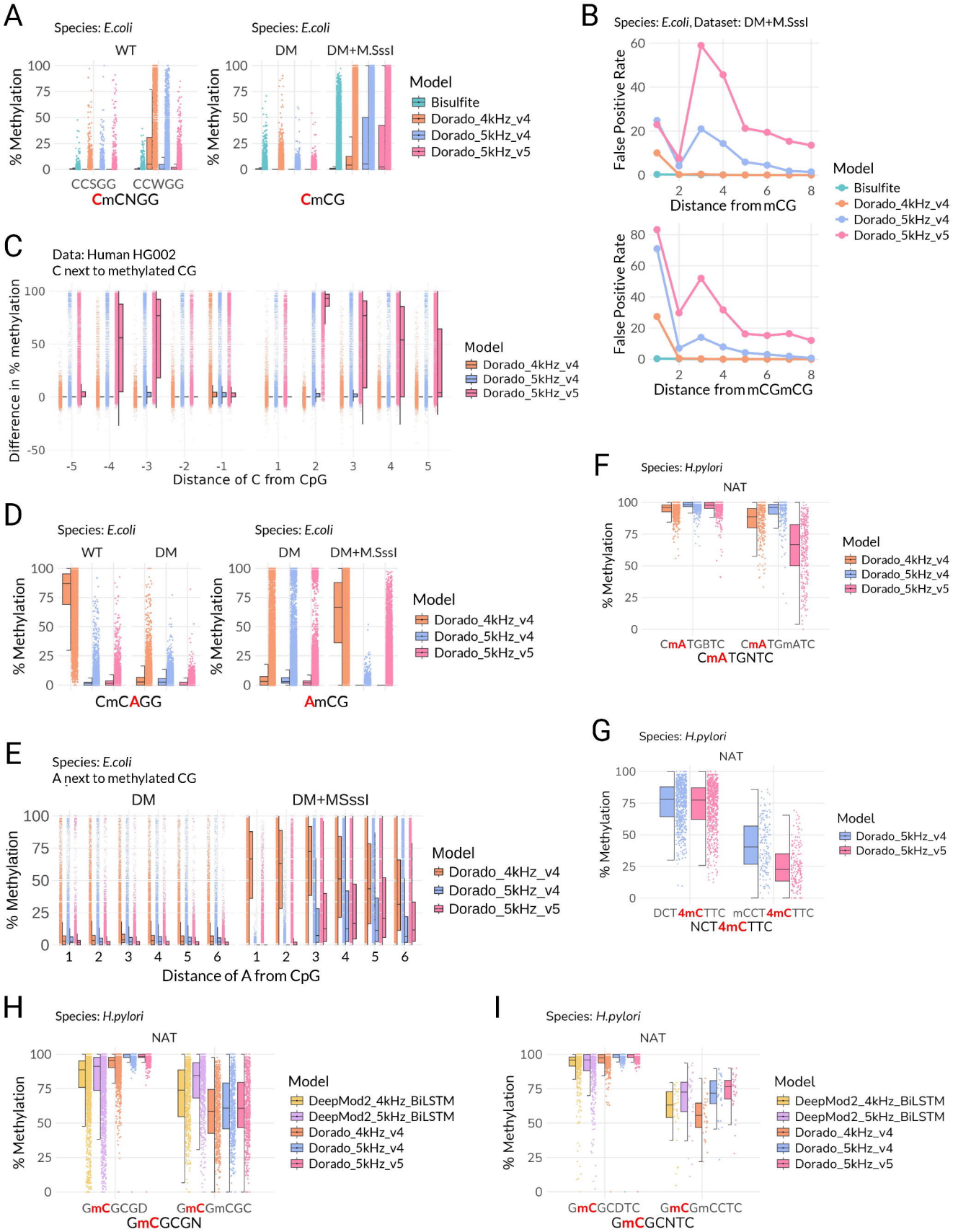

A
Species: E.coli | Motif: mCG
F1 score
Coverage
Model
DeepMod2_4kHz_BiLSTM
Dorado_4kHz_v4
DeepMod2_5kHz_BiLSTM
Dorado_5kHz_v4
Dorado_5kHz_v5
B
Species: E.coli | Motif: CmCWGG
Dorado_4kHz_v4
Dorado_5kHz_v4
Dorado_5kHz_v5
C
Species: E.coli | Motif: GmATC
Species: H.pylori | Motif: GmATC
D
Species: H.pylori
ATTAmAT
CmATG
GmANTC
E
Effect of minimum read quality | Species: E.coli
GmATC
CmCWGG
Score
Minimum read quality
F1
Recall
Precision
Specificity
mCG
DeepMod2_4kHz
DeepMod2_5kHz

Figure 5


A
Species: E.coli
WT
DM
DM+M.SssI
% Methylation
CCSGG
CCWGG
CmCNGG
CmCG
Model
Bisulfite
Dorado_4kHz_v4
Dorado_5kHz_v4
Dorado_5kHz_v5
B
Species: E.coli, Dataset: DM+M.SssI
False Positive Rate
Distance from mCG
Distance from mCGmCG
C
Data: Human HG002
C next to methylated CG
Difference in % methylation
Distance of C from CpG
D
CmCAGG
AmCG
E
Species: E.coli
A next to methylated CG
DM+MSssI
Distance of A from CpG
F
Species: H.pylori
NAT
CmATGBTC
CmATGmATC
CmATGNTC
G
DCT4mCTTC
mCCT4mCTTC
NCT4mCTTC
H
DeepMod2_4kHz_BiLSTM
DeepMod2_5kHz_BiLSTM
GmCGCGD
GmCGmCGC
GmCGCGN
I
GmCGCDTC
GmCGmCCTC
GmCGCNTC