Title

Mapping DNA methylation with high-throughput nanopore sequencing

Permalink

https://escholarship.org/uc/item/0098k5s0

Journal

Nature Methods, 14(4)

ISSN

1548-7091

Authors

Rand, Arthur C
Jain, Miten
Eizenga, Jordan M
et al.

Publication Date

2017-04-01

DOI

10.1038/nmeth.4189

Peer reviewed

Published in final edited form as:
Nat Methods. 2017 April ; 14(4): 411–413. doi:10.1038/nmeth.4189.

Mapping DNA Methylation with High Throughput Nanopore Sequencing

Arthur C. Rand[*], **Miten Jain**[*], **Jordan M. Eizenga**[*], **Audrey Musselman-Brown**, **Hugh E. Olsen**, **Mark Akeson**, and **Benedict Paten**
Department of Biomolecular Engineering, University of California, Santa Cruz

Genomics Institute, University of California, Santa Cruz

Abstract

Chemical modifications to DNA regulate its biological function. We present a framework for mapping methylation to cytosine and adenosine with the Oxford Nanopore Technologies MinION using its ionic current signal. We map three cytosine variants and two adenine variants. The results show that our model is sensitive enough to detect changes in genomic DNA methylation levels as a function of growth phase in *E. coli*.

Chemical modifications of DNA influence its biological function. In mammals, alkylation of carbon-5 yields several cytosine variants: C^5-methylcytosine (5-mC), C^5-hydroxymethylcytosine (5-hmC), C^5-formylcytosine, and C^5-carboxylcytosine. These marks play a role in aging, gene regulation, imprinting, and disease[1,2]. In prokaryotes, 5-mC and N^4-methylcytosine contribute to gene regulation and restriction-modification systems[3]. Some eukaryotic and prokaryotic organisms also methylate adenine to N^6-methyladenine (6-mA). This residue is important for a variety of biological processes including methylation-dependent mismatch repair and transcriptional regulation[3–7].

Several methods can detect DNA methylation. Illumina-based sequencing uses bisulfite treatment to detect cytosine methylation. Single-molecule real time (SMRT) sequencing generates long reads and can detect multiple modifications to DNA simultaneously using enzyme kinetics[8–14]. Previous studies have shown that ionic current measurements from low-throughput nanopore sensors can discriminate among all five C5-cytosine variants[15,16].

Here we show that DNA modifications can be detected as changes in the Oxford Nanopore Technologies (ONT) MinION's ionic current signal. The MinION is a high-throughput

Correspondence to: Benedict Paten.

[*] These authors contributed equally to this work.

Author contributions
B.P. conceived of the experiments. B.P. and M.A. directed the research. A.C.R. implemented the models and performed analysis. M.J. and H.E.O. performed the sequencing experiments and performed sequence data analysis. J.M.E. implemented the HDP model and Gibbs sampler. A.M-B. performed initial experiments. All authors contributed to writing the manuscript.

Competing financial interests
M.A. is a consultant to Oxford Nanopore Technologies.

nanopore-based single-molecule device that can sequence long unamplified DNA fragments[17]. We present a generative model that can perform inference on the methylation status of individual bases in a reference sequence. The model consists of a variable-order hidden Markov model (HMM) with a hierarchical Dirichlet process (HDP) to learn ionic current distributions (Figure S1, **Online Methods**). We refer to the full model as an HMM-HDP. We classify individual C, 5-mC, and 5-hmC bases on single molecules of synthetic oligonucleotides. We also map 5-mC at the inner cytosine of C**C**(A/T)GG motifs and 6-mA at G**A**TC motifs in *E. coli* genomic DNA (gDNA) from multiple growth conditions to demonstrate that we can quantify changes in methylation levels in realistic conditions.

The MinION continuously records ionic current and then divides it into segments referred to as events (See **Online Methods** for details). Our method models each event as a nucleotide string of length *k*, a k-mer. Each k-mer is associated with a distribution of ionic currents in picoamps (pA). We model the events with a pair-HMM that tracks a reference sequence and allows reference nucleotides to be any of several potentially modified bases (Figure S1A and B). We use a HDP mixture model to learn the effects that different base modifications have on the ionic current (Figure S1C and D), the HDP ties together the parameters between different k-mer distributions[18]. Our method calls methylation variants based on the posterior probability of event-to-k-mer aligned pairs. We call the methylation status as the variant with the highest marginal probability. With multiple reads, we sum the probabilities from the individual reads aligned to a position and call the variant with the highest posterior mean. Details about the model can be found in the **Online Methods**.

We evaluated our method's performance at discriminating cytosine variants with a three-way cytosine classification experiment with synthetic oligonucleotide substrates, where each molecule bears only C, 5-mC, or 5-hmC. See **Online Methods** for details. We also used these experiments to compare our model to a more naive HMM in which emission distributions are maximum likelihood normal distributions. The nucleotide sequences required the model to learn 2,868 new 6-mer ionic current distributions with methylated bases in addition to the 1,784 canonical ones. We measured the per-read accuracy by the proportion of correct methylation calls on a single strand. The template reads had higher classification accuracy (74% and 80% mean and median accuracy, respectively) than the complement reads (67% and 76% mean and median accuracy, respectively), corroborating previous results on average sequence identity (Figure 1A)[19]. The HMM-HDP performance was also significantly better than the simpler HMM(Table S1). The HMM-HDP model classified different cytosines at accuracies ranging from 16% to 95% with median accuracy of 76% for template reads and 70% for the complement reads (Figure 1B, Table S1). Calling sites as unmodified cytosine is the most common error, with 5-mC being the most commonly miscalled variant (Figure 1C, S2).

We hypothesized that variability in sequence context accuracy results from ionic current distributions that vary only slightly between the methylation states. To test this hypothesis, we compared the mean pairwise Hellinger distance between the ionic current distributions of the 6-mers overlapping a site and the site's classification accuracy (Figure 1D). The Pearson correlation was 0.52 (p = 6.6E-33, t = 12.98, df = 445) on the template strand and 0.36 (p = 9.0E-15, t = 8.02, df = 445) on the complement strand (Figure 2D, S3), suggesting that there is indeed a relationship between the similarity of the distributions and methylation calling accuracy.

To test our method's ability to map cytosine and adenine methylation in a controlled system, we sequenced pUC19 plasmid DNA grown in *E. coli* containing both *dam* and *dcm* methyltransferases. These substrates are completely methylated at C**C**(A/T)GG and G**A**TC motifs[4]. The mean per-read cytosine variant calling accuracy was 79% and 72% on the template and complement strands, respectively. The accuracy for calling adenine variants on the two strands was 70% and 58%. The ionic current of events mapped to cytosine motifs showed a more pronounced difference between methylation states than events mapped to adenine motifs (Figure 2, top), likely contributing to the lower accuracy calling adenine variants.

To assess the effect of variations in data quality among reads, we explored the relationship between the accuracy and the ungapped alignment score, a proxy for data variation. The variation in per-read accuracy was correlated with the ungapped alignment score (Figure 2-bottom). To assess our method's ability to call methylation variants with multiple reads, we randomly sampled 40× coverage and called cytosine and adenine variants. The best cutoff values classified 96% and 86% of the residues correctly for cytosine and adenine residues, respectively (Table S2A and B).

To evaluate the model's ability to classify 5-mC in a more realistic experimental setup, we mapped 5-mC at C**C**(A/T)GG motifs in *E. coli* gDNA and PCR-amplified DNA (pcrDNA). We evenly divided 3,418 constitutively methylated cytosines into a training and testing set. We assumed that none of cytosines in the DNA amplicons were methylated and that all of the cytosines in the gDNA were methylated. Based on these labels, the model was able to correctly classify 96% of the cytosines motifs in the test set (Table S2C).

Chemical modifications to DNA happen post-replication, and they often depend on the state of the cell. One condition that is known to affect methylation levels in *E. coli* is growth phase[4]. We sequenced gDNA isolated from *E. coli* cultures harvested at three different growth phases: early-exponential (0.4 OD), late-exponential (0.8 OD), and stationary (24 hours). For cytosine classification, we used both template and complement reads and called 23,004 (95.5%), 23,789 (98.7%), and 24,034 (99.8%) of the cytosines as methylated in the early-exponential, late-exponential, and stationary-growth phases respectively. These results are consistent with previous studies that showed increasing levels of cytosine methylation from early-exponential phase growth through stationary phase growth (Figure 3, Table S3)[20].

In the adenine classification experiments, we only used template reads to assay for methylation levels because our previous experiments with pUC19 plasmid DNA showed that using only template reads gave the highest accuracy (Table S2A). The classifier called 33,930 (89%), 34,884 (91%), and 31,901 (83%) of the adenines as methylated in the early-exponential, late-exponential, and stationary growth phases respectively. Transcriptional levels of *dam* have been shown to reach a maximum during exponential phase growth, followed by a decrease during stationary phase growth[5,21]. Our results are consistent with this pattern (Figure 3, Table S3). Details of how the models were evaluated are described in the **Online Methods**.

Using the MinION's ionic current signal, we achieved a median three-way cytosine methylation classification read accuracy of 80% on synthetic DNA. We tested our method in a model system by mapping 5-mC and 6-mA bases in *E. coli* gDNA and plasmid gDNA. We correctly mapped the methylation status of 96% of the cytosines in E. coli DNA and 86% of the adenines in pUC19 plasmid DNA with 20× and 40× coverage, respectively. To demonstrate the utility of the method in a dynamic system, we show that genome-wide changes in methylation at different growth phases in *E. coli* can be detected even with imperfect training data.

We anticipate numerous applications for this method. For instance, it could be used to phase multiple base modifications simultaneously on long reads. In addition, since our method does not require any additional sample preparation, this information is available in any MinION sequencing experiment that uses gDNA. Lastly, changing the set of base modifications our model detects is straightforward as long as there is appropriate training data (Supplementary Discussion). We intend to develop and release models that detect a broader array of modifications in the future.

Online Methods

MinION sequencing

The sequencing runs on synthetic oligonucleotides were performed in late 2015 using R7.3 chemistry (SQK-MAP006 sequencing kits). The R7.3 MinION sequencing protocol records ionic current at 3kHz and modeled event as corresponding to 6-mers. The pUC19 plasmid DNA, *E. coli* gDNA and pcrDNA were sequenced using R9 chemistry (EXP-NSK007 sequencing kits). The R9 version uses a different pore and increased sequencing speed. In this version of the protocol, the MinION samples ionic current at 4kHz and the events are modeled as 5-mers. We initially used a 6-mer lookup table for the R9 pore provided by ONT, then estimated our own 5-mer model from a collection of reads (Supplemental Information - Estimating emission distributions for R9 nanopores).

Sequencing controlled synthetic DNA substrates containing C, 5-mC, or 5-hmC

We used 897 bp synthetic DNA oligonucleotides from ZYMO Research (Catalog # D5405) that contain entirely C, 5-mC, or 5-hmC bases. Apart from the cytosines, the oligonucleotides have identical sequences. We performed sequencing experiments using R7.3 chemistry (SQK-MAP006 sequencing kits) with four MinION flow cells: one for each of the three substrates, and one where all the substrates were with barcoded with uniquely identifying sequences (EXP-NBD001 barcoding kit) and run together on one flow cell. The runs where the strands were sequenced individually produced 68,920, 27,073, and 70,641, reads for the C, 5-mC, and 5-hmC strands, respectively. The run where the strands were barcoded and sequenced together produced 6,966, 294, and 467 reads for the C, 5-mC, and 5-hmC strands, respectively. The reads spanned the full length of the substrate. All models were trained on the reads where the strands were run in separate flow cells. The bar-coded reads served as our test dataset. This experimental design maximized the amount of training data while controlling for batch effects between MinION runs. Sequence data were processed using Metrichor (versions 1.15.0 and 1.19.0), and only "pass" 2D reads that covered the full length of the reference sequence were used for downstream analysis

Preparation of DNA control substrates containing 6-mA and 5-mC

We purchased pUC19 vector DNA from New England Biolabs (NEB cat. number N3041S). This DNA is isolated from *E. coli* strain ER2272 that contains genes methyltransferase (MTase) genes *dam* and *dcm*. The *dam* MTase methylates the adenine in G**A**TC sequence contexts and the *dcm* MTase methylates the inner cytosine at C**C**(A/T)GG sequence contexts (bold letter indicating methylated position). We linearized the plasmid by restriction digest at a unique SspI (NEB cat. number R0132S) restriction site. The linearized plasmid was purified by excising the band from an agarose gel following electrophoresis. The DNA was eluted from the gel using the Wizard SV kit (Promega) as per manufacturer's instructions. To generate an unmethylated substrate, we PCR amplified the plasmid with primers around the SspI restriction site (forward: 5′ ATT ATT GAA GCA TTT ATC AGG GTT ATT GTC, reverse: 5′ ATT GAA AAA GGA AGA GTA TGA GTA TTC AAC) with Q5 high-fidelity polymerase master mix (NEB cat. number M0492S) as per the manufacturer's specifications. The PCR reaction was purified with 0.4× AMPure SPRI beads using standard procedures.

Sequencing for pUC19 plasmid DNA

We sequenced the methylated plasmid and an unmethylated PCR amplicon in the same flow cell (see Methods). The pUC19 DNA sequence is 2,686 bp long. It contains 30 adenine residues in GATC motifs and 10 cytosines at CC(A/T)GG motifs. The motifs are palindromic, so they each contain two potentially modified residues: one on each strand. The reads covered the entire length of the substrate.

The purified PCR-amplified and linear pUC19 DNA were individually barcoded (EXP-NBD002 barcoding kit). Roughly equimolar amounts of the barcoded material was combined and sequenced on the MinION using R9 chemistry (NSK-007 sequencing kit). The sequencing run produced 27,293 and 17,220 pass 2D reads in the pcrDNA and gDNA, respectively.

Data selection and partitioning for experiments with pUC19 DNA

Reads that covered the entire length of the plasmid sequence were shuffled and 40% of the reads were used to train the models and the remaining 60% for testing (see *Supervised training of model parameters*). For methylation variant calling, 40 full-length reads were randomly selected from the test read alignments and used to call the methylation status of the adenines and cytosines in the reference. This process was iterated 100 times to generate an error rate distribution.

Sequencing for genomic and amplified *E. coli* DNA

We performed one sequencing run using standard procedures on genomic gDNA from *E. coli* strain K-12 MG1655 (DSMZ) and another run on DNA that was pcrDNA using a whole-genome amplification kit (Qiagen REPLI-g). Both runs were done independently using R9 chemistry (EXP-NSK007 sequencing kits). The gDNA run produced 18,177 pass 2D reads (132 Mb) with an average read length of 7.3 kb. The pcrDNA run produced 61,408 pass 2D reads (387 Mb) with an average read length of 6.3 kb. The reads were shuffled and evenly divided into two groups, one was used for training the model and the other for classification experiments.

Data selection and partitioning for experiments with *E. coli* DNA

Previous research has mapped the locations of 5-mC on the CC(A/T)GG motifs using bisulfite sequencing[2020]. These data reported 1,709 high-confidence methylated motifs in stationary phase cells. We divided the motifs into a training group and test group. Care was taken to be sure that k-mers in the test group were observed in the training group (Supplementary Methods - Dividing *E. coli* methylation motifs into training and test groups). The HDP-HMM was trained on alignments generated with pcrDNA reads supplemented with events from gDNA reads that aligned to the high-confidence methylated sites from the training group. We used the trained model to classify the methylation status of cytosines in the test group motifs from the held out portion of reads from the pcrDNA and gDNA sequencing runs.

The models used in the growth phase experiments were trained it on all 3,418 known-methylated cytosines. We trained the model on reads from stationary phase genomic DNA and PCR amplified DNA. We evaluated the model by classifying the 3,418 known cytosines, accuracy and precision were 96% and 92%, respectively. To train the adenine classification model we labeled all adenines at G**A**TC sites in the *E. coli* genome as methylated in reads from stationary phase cells (see Supplemental Information - Adenine classification with approximate labels). In total, we classified 24,100 cytosines at C**C**(A/T)GG motifs and 38,248 adenines at G**A**TC motifs. To directly evaluate the accuracy of the model we called variants on the pUC19 plasmid using the procedure described previously. The model had an estimated accuracy and precision of 87(+/−)3% and 84(+/−)4%, respectively. However, the pUC19 sequence does not contain all of the GATC contexts in the *E. coli* genome, so this measure of accuracy may not fully generalize. Our results are also concordant with previously described results using SMRT sequencing (see Supplementary Discussion).

Hierarchical Dirichlet process mixture model

The HDP mixture is a statistical model in which a collection of mixture distributions are composed of a countably infinite set of shared mixture components. The weights of the components in each mixture distribution are determined according to a separate Dirichlet process on the shared collection of components[19]. In addition, the mixture components themselves are distributed according to a Dirichlet process that draws components from a base distribution (Supplementary Methods - Hierarchical Dirichlet Process Mixture Model for Ionic Current Distributions). In our model, the base distribution is the normal-inverse gamma distribution, which is a conjugate prior to the normal distribution (that is, to the mixture components).

Sharing mixture components statistically shrinks our estimates of the current distributions toward each other (Supplementary Methods - Grouping 6-mers with different HDP topologies). This boosts statistical strength since each distribution can share the information learned by the others. We also have the option of adding a further layer of Dirichlet processes between the Dirichlet process that generates the distribution over shared components and the Dirichlet processes that generate the k-mer distributions (Figure S1D). This encourages a greater degree of shrinkage within each subtree. We experimented with several topologies for this tree, each representing a different grouping of k-mers based on their sequence composition (see *Structure of the hierarchical Dirichlet process*).

Structure of variable-order hidden Markov model

Our HMM is structured to allow alignment of multiple different bases at a given position in the reference sequence. We term these positions *ambiguous positions*. Positions in the reference are designated ambiguous before the alignment begins. In three-way classification experiments on synthetic oligonucleotides, we allow for C, 5-mC, and 5-hmC to be aligned to a given cytosine. In two-way classification experiments, the model is restricted to C and 5-mC or A and 6-mA in the cases of alignment to cytosine and adenine, respectively. These experiments also restrict ambiguous sites to the known methylation motifs.

The fact that each event corresponds to multiple positions in the reference means that more than one event reports on a single ambiguous position. Accordingly, we tie the probabilities of consistent methylation variant calls by configuring our HMM in a variable-order meta-structure that allows for multiple paths over a reference k-mer depending on the number of methylation possibilities (Supplementary Methods - Variable-Order Hidden Markov Model). The dynamic programming matrix has high-dimensional cells to accommodate these paths. We restrict the recursion by only allowing transitions if the bases in positions 2-k in the first k-mer are identical to the bases at positions 1-(k-1) in the second k-mer (Figure S1B, Figure S5). The joint probability for the read's event sequence and the reference is calculated with the forward-backward algorithm, and the likelihood of each methylation variant at each ambiguous position is calculated by marginalizing over the HMM's states. We treat the template and complement event sequences as independent, so a given event sequence is aligned to the appropriate nucleotide sequence and reports on only the cytosines in that strand. To make methylation variant calls using multiple reads, the variant with the greatest posterior mean given the reads is called.

Structure of the hierarchical Dirichlet process

The HDP bolsters its statistical strength by sharing information between the set of distributions it estimates. In effect, this encourages the distributions to be more similar to each other than if they were modeled independently. The HDP model also can encourage a greater degree of similarity between pre-specified subgroups of distributions (see Methods and Supplemental Information - Grouping 6-mers with different HDP topologies). This can increase statistical strength further, assuming that the subgroups reflect clusters of similarity in the true distributions.

Since the biophysical relationship between each given k-mer sequence and the observed ionic current distribution is poorly understood, we empirically tested whether different subgroupings would increase statistical strength using reads from the synthetic oligonucleotides. We tested HDP models with five different subgroupings of 6-mers. The two-level HDP does not separate them into any subgroups (Figure S1C), whereas the rest of the models group 6-mers by features of their 6-mer sequence (Figure S1D). The "Multiset" HDP. The "Composition" HDP groups 6-mers by how many purines and pyrimidines they contain. The "MiddleNucleotides" HDP groups 6-mers based on the center two bases in the 6-mer. Finally, the "GroupMultiset" HDP groups the 6-mers by their nucleotide content without regard to their order or their methylation status. The best performing model was the "Multiset" model (Table S1, **Online Methods**). Although it was a small gain in accuracy over the simpler ungrouped model, we used the "Multiset" model for all further analyses.

Mapping of reads and event alignment

We align the ionic current events from each read to the reference sequence in a two step process. First we generate a guide alignment between the read's nucleotide sequence and the reference, which we then use to constrain a second alignment of events to the reference. The guide alignment uses a concatenated sequence from Metrichor's '2D alignment' table, which allows for each base to be mapped to an event in the template and complement event sequence (Supplementary Methods - Making a read sequence from the 2D alignment table). This nucleotide sequence is aligned to the reference with BWA-MEM in ont2d mode[22]. Runs of consecutive matches in the guide alignment serve as anchors for the event sequence alignment. The anchors are mapped back to events in the template and complement event sequences. The dynamic programming is constrained by these anchors similar to the method described by Paten et al.[23] (Supplementary Methods - Banded alignment). The template and complement event sequences are independently aligned to the reference using the HMM described below.

Computing posterior probabilities of alignments and ungapped alignment scores

Our HMM takes the event sequences as input and aligns them to a reference nucleotide sequence. Let x_i be a k-mer in the reference sequence S, and e_j be an event in the event sequence E, and $e_j x_i$ mean that event e_j is aligned to k-mer x_i. The model calculates $P(e_j x_i| E, S, \Theta)$, the posterior probability for event/k-mer aligned pairs given the model Θ. We also leverage the posterior probabilities to compute a measure of alignment quality, the ungapped alignment score, U.A.S., which is defined as

$$\frac{\sum_{(e_j, x_i) \in \pi} P(e_j \Diamond x_i | E, S, \Theta)}{N}$$

where N is the total number of aligned pairs in the alignment π.

Generating preliminary alignments without consideration for methylation status

ONT provides a lookup table of parameters describing normal distributions that they use characterizes the current distributions of the 4096 canonical base 6-mers both for R7.3 and R9 sequencing protocols. In the case of R9, however, we estimated a new 5-mer lookup table for the 1,024 canonical k-mers (Supplemental Information - Estimating emission distributions for R9 nanopores). We leverage this table to heuristically initialize the emission distributions in our HMM-HDP over the expanded alphabet. To do so, we generate a preliminary alignments using the lookup table and then infer the methylation status of the events based on their flow cell or barcode. We can then use high probability aligned pairs from this alignment to train the emission distributions of the HMM-HDP, for more details see Supplemental Information - Training the HMM-HDP model.

Supervised training of model parameters

We train the HMM with a variant of the Baum-Welch procedure. First, we heuristically initialize the emission distributions by training them on aligned events above a probability threshold (0.9 for the synthetic oligonucleotides and 0.8 for the *E. coli* and plasmid DNA) from the preliminary alignment described above. In the three-way classification control experiments on synthetic oligonucleotides using normal distributions, this entails calculating the maximum likelihood normal distribution for each 6-mer. For the HMM-HDP, we estimate the posterior mean density for each k-mer's distribution using a Markov chain Monte Carlo (MCMC) algorithm (Supplementary Methods - Training the HMM-HDP model). In both cases, only the distributions for the ionic current means are learned following the preliminary alignment. A separate neural net experiment suggested that the event noise did not add to classification accuracy (Supplementary Methods - Classification of Ionic Current Events with Neural Networks). At this step, we also re-estimate the HMM's transition probabilities independently. We then produce new alignments and re-estimate the emission distributions from high confidence assignments as in the initialization. This process is iterated until the model's variant calling accuracy stops improving.

The MCMC algorithm we use for the HDP is the Chinese Restaurant Franchise Algorithm, a Gibbs sampler for HDP mixture models[1818]. We discard a number of burn-in iterations equal to 30-times the total number of assignment data points and then collect 10,000 samples, thinning sampling iterations by 100. Whenever we record samples from the Markov chain, we evaluate the posterior predictive distribution for each 6-mer at a grid of 1200 evenly spaced points in the interval between 30 pA and 90 pA for R7.3. For R9 experiments we use a grid of 1800 evenly spaced points in the interval between 50 pA and 140 pA. After sampling, we compute our estimate of the posterior mean density as the mean of the sampled densities at each grid point. Subsequently, we interpolate within the grid using natural cubic splines.

Code availability

All relevant code can be accessed at: https://github.com/ArtRand/signalAlign. Code for reproducing results and examples of how to use the described model can be found at: https://github.com/ArtRand/CytosineMethylationAnalysis.

Data availability

All generated *E. coli*, pUC19, and synthetic oligo data is available upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research reported in this publication was supported by the National Human Genome Research Institute of the US National Institutes of Health under award numbers HG006321 (M.A.), HG007827 (M.A.) and 5U54HG007990 (B.P.).

References

1. Schübeler D. Function and information content of DNA methylation. Nature. 2015; 517:321–326. [PubMed: 25592537]
2. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat. Rev. Genet. 2013; 14:204–220. [PubMed: 23400093]
3. Sánchez-Romero MA, Cota I, Casadesús J. DNA methylation in bacteria: from the methyl group to the methylome. Curr. Opin. Microbiol. 2015; 25:9–16. [PubMed: 25818841]
4. Marinus MG, Løbner-Olesen A. DNA Methylation. EcoSal Plus. 2014; 6
5. Marinus MG, Casadesus J. Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more. FEMS Microbiol. Rev. 2009; 33:488–503. [PubMed: 19175412]
6. Li J, et al. Epigenetic Switch Driven by DNA Inversions Dictates Phase Variation in Streptococcus pneumoniae. PLoS Pathog. 2016; 12:e1005762. [PubMed: 27427949]
7. Greer EL, et al. DNA Methylation on N6-Adenine in C. elegans. Cell. 2015; 161:868–878. [PubMed: 25936839]
8. Flusberg BA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat. Methods. 2010; 7:461–465. [PubMed: 20453866]
9. Frommer M, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc. Natl. Acad. Sci. U. S. A. 1992; 89:1827–1831. [PubMed: 1542678]
10. Davis BM, Chao MC, Waldor MK. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. Curr. Opin. Microbiol. 2013; 16:192–198. [PubMed: 23434113]
11. Cohen NR, et al. A role for the bacterial GATC methylome in antibiotic stress survival. Nat. Genet. 2016; doi: 10.1038/ng.3530
12. Beaulaurier J, et al. Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. Nat. Commun. 2015; 6:7438. [PubMed: 26074426]
13. Booth MJ, et al. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nat. Protoc. 2013; 8:1841–1851. [PubMed: 24008380]
14. Saletore Y, et al. The birth of the Epitranscriptome: deciphering the function of RNA modifications. Genome Biol. 2012; 13:175. [PubMed: 23113984]
15. Wescoe ZL, Schreiber J, Akeson M. Nanopores discriminate among five C5-cytosine variants in DNA. J. Am. Chem. Soc. 2014; 136:16582–16587. [PubMed: 25347819]
16. Laszlo AH, et al. Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. Proc. Natl. Acad. Sci. U. S. A. 2013; 110:18904–18909. [PubMed: 24167255]
17. Ip CLC, et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Res. 2015; 4:1075. [PubMed: 26834992]
18. Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet Processes. J. Am. Stat. Assoc. 2006; 101:1566–1581.
19. Jain M, et al. Improved data analysis for the MinION nanopore sequencer. Nat. Methods. 2015; 12:351–356. [PubMed: 25686389]
20. Kahramanoglou C, et al. Genomics of DNA cytosine methylation in Escherichia coli reveals its role in stationary phase transcription. Nat. Commun. 2012; 3:886. [PubMed: 22673913]
21. Seshasayee ASN. An Assessment of the Role of DNA Adenine Methyltransferase on Gene Expression Regulation in E coli. PLoS One. 2007; 2:e273. [PubMed: 17342207]
22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

23. Paten B, Herrero J, Beal K, Birney E. Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. Bioinformatics. 2009; 25:295–301. [PubMed: 19056777]

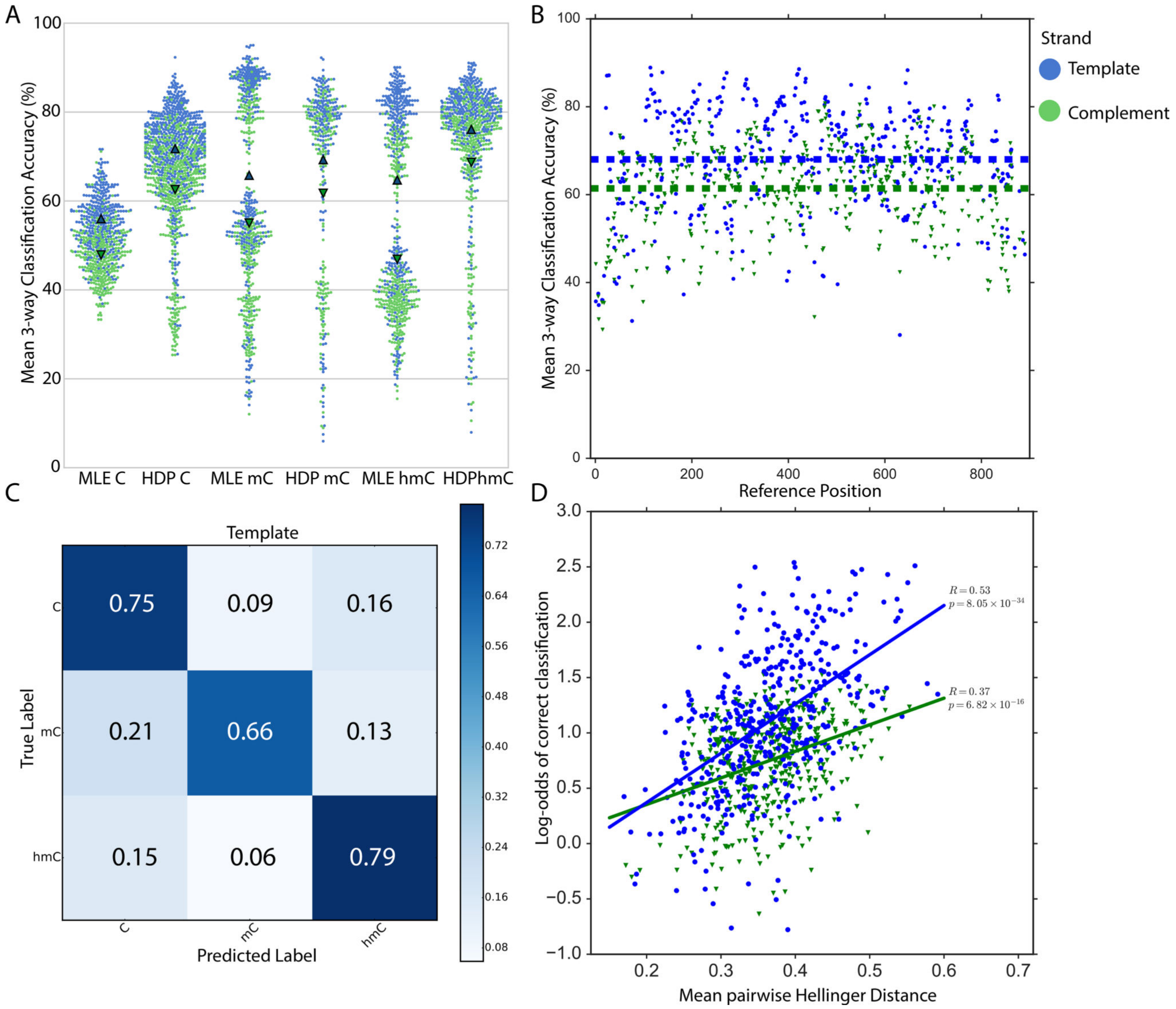

Figure 1.

Cytosine methylation variant calling accuracy results on synthetic oligonucleotides. Results are from classification of 6,966, 294, and 467, C, 5-mC, and 5-hmC strands respectively that were barcoded and sequenced in the same MinION flow cell. A. Per-read accuracy distribution is shown for the maximum-likelihood estimate (MLE) normal distributions and the ‘Multiset’ HDP model. The triangles represent the mean of the distribution. B. Average three-way classification accuracy for all sites on the substrate. Dotted lines represent the mean across all sites for template (blue) and complement reads (green). C. Confusion matrix showing HMM-HDP three-way cytosine classification performance on template reads of synthetic oligonucleotides. D. Scatter plot showing the correlation between the log-odds of correct classification and the mean pairwise Hellinger distance between the methylation statuses of the 6-mer distributions overlapping a cytosine.

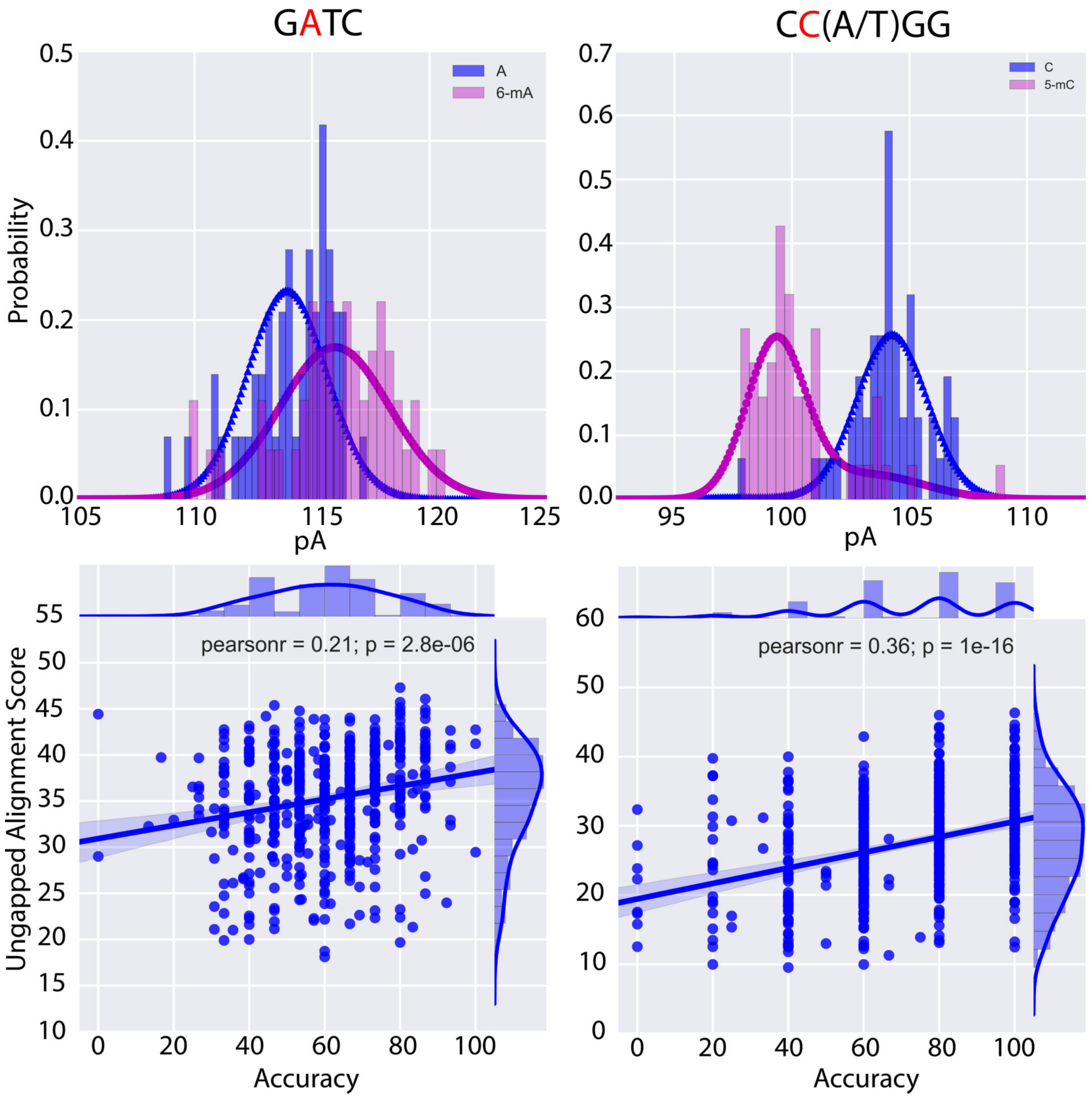

Figure 2.

Observed and learned ionic current distributions and read accuracy correlation with ungapped alignment score for 6-mA in GATC motifs (left) and 5-mC in CC(A/T)GG motifs (right). Top: Comparison of the influence of 6-mA and 5-mC on ionic current levels for representative 5-mers. The empirical ionic current levels from 100 aligned events are shown as a normalized histogram and the HDP-learned probability densities were are shown as curves. The HDP density was sampled on 900 point grid from 50 to 140 pA. Bottom. Correlation between ungapped alignment score (see Methods) and per-read accuracy for 500 randomly sampled template reads.

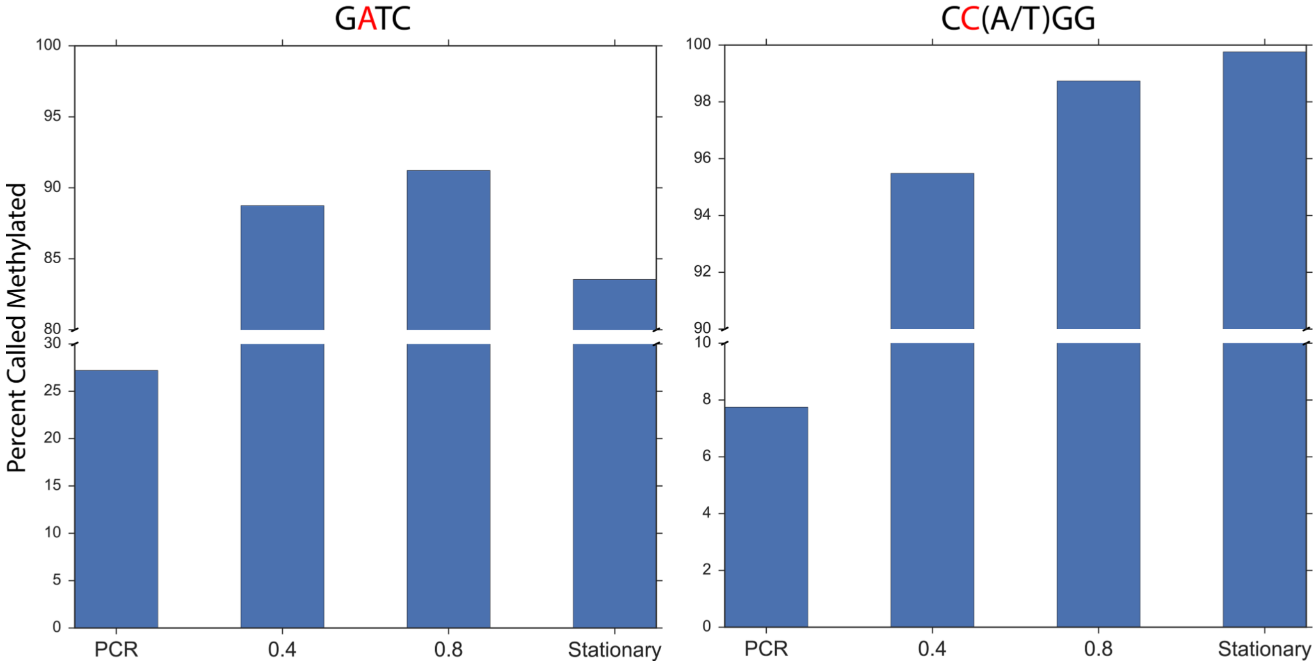

Figure 3.

Changes in genome-wide cytosine methylation at different stages of culture growth. Bar height represents the percentage of residues that were called as methylated. Axes are broken to accentuate differences between the growth phases.