Epigenetic Exploration with Nanopore Sequencing

Isac Lee[1], Rachael Workman[1], Yunfan Fan[1], Norah Sadowski[1], Jared Simpson[2], Winston Timp[1]

[1]Johns Hopkins University Department of Biomedical Engineering
[2]Ontario Institute for Cancer Research

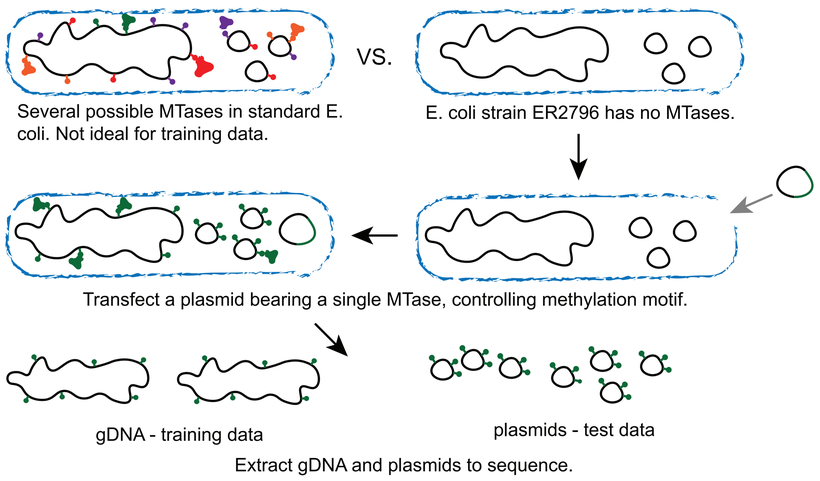
Abstract

Nanopore sequencing has enormous potential in epigenetic applications; unlike traditional sequencing-by-synthesis technologies, it can distinguish covalently modified nucleotides directly through their modulation of the electrolytic current. We can take advantage of the long read lengths (>10kb) generated by nanopore sequencing to precisely call methylation patterns, and to obtain phased methylation information across the genome. We have already demonstrated the accuracy and some of the promise of nanopore sequencing in calling 5-methylcytosine in a CG context using a hidden Markov model(Simpson et. al *Nature Methods* 2017). We have now trained our model on other modified (non-canonical) nucleotides, and in other contexts, including N6-methyladenosine (N6-mA), 4-methylcytosine, and 5-methylcytosine in a non-CG context.

Modifications have applications in bacterial and mammalian epigenetics, but we may be able to leverage unnatural modifications to label different aspects of the epigenome. A good example of this is the NOMe-seq (Kelly, *Genome Res.* 2012) method, which uses GpC methylation to label accessible, open chromatin. We have demonstrated this application, simultaneously detecting GpC and CpG methylation on the same molecule, providing two distinct epigenetic signatures along ~10kb single DNA molecules.

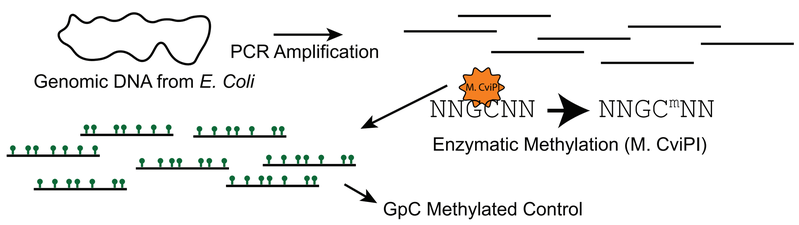
Methods

Training 4mC and 6mA - native methylation



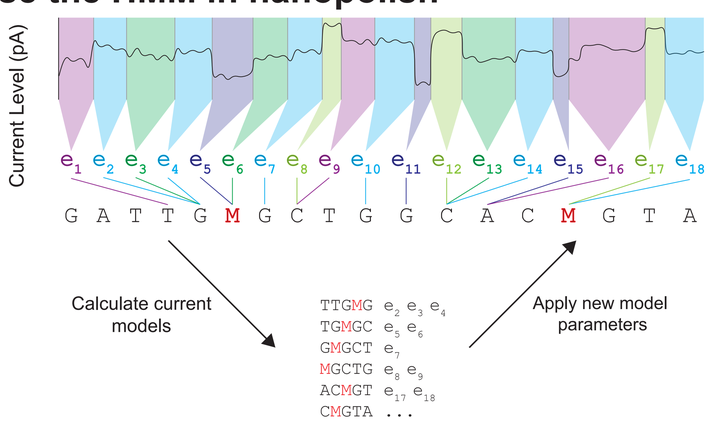
Through our collaborators at NEB, we have obtained gDNA and plasmids encoding for different methyltransferase proteins - including proteins which generate 4-methylcytosine, 5-methylcytosine and N6-methyladenosine. These methyltransferases are transformed into a methylation transferase negative bacteria (ER2796), where the expressed methyltransferase acts to label the gDNA and plasmid DNA. The gDNA and plasmids are then extracted and we subsequently generated nanopore libraries via shearing, end-polishing and sequencing adapter ligation prep. These libraries were then sequenced on minIONs with v9.4 flowcells.

Training Gp^{5m}C - Enzymatic methylation



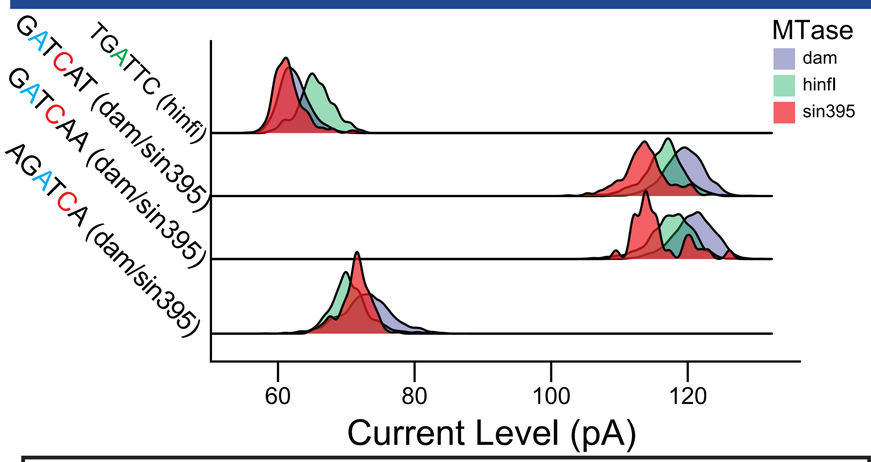
Alternatively, when a purified enzyme cocktail is available, we use that to perform an *in vitro* modification reaction. First, we generate a completely unmethylated control sample by shearing E. Coli gDNA, ligating universal adapters and amplifying. We then treat a portion of this with M. CviPI to convert GC to GCm. We previously (Simpson et al.) generated such a training set using M. Sssi (CpG methyltransferase) - shown here is a schematic of our new training set using M. CviPI (GpC methyltransferase)

Use the HMM in nanopolish

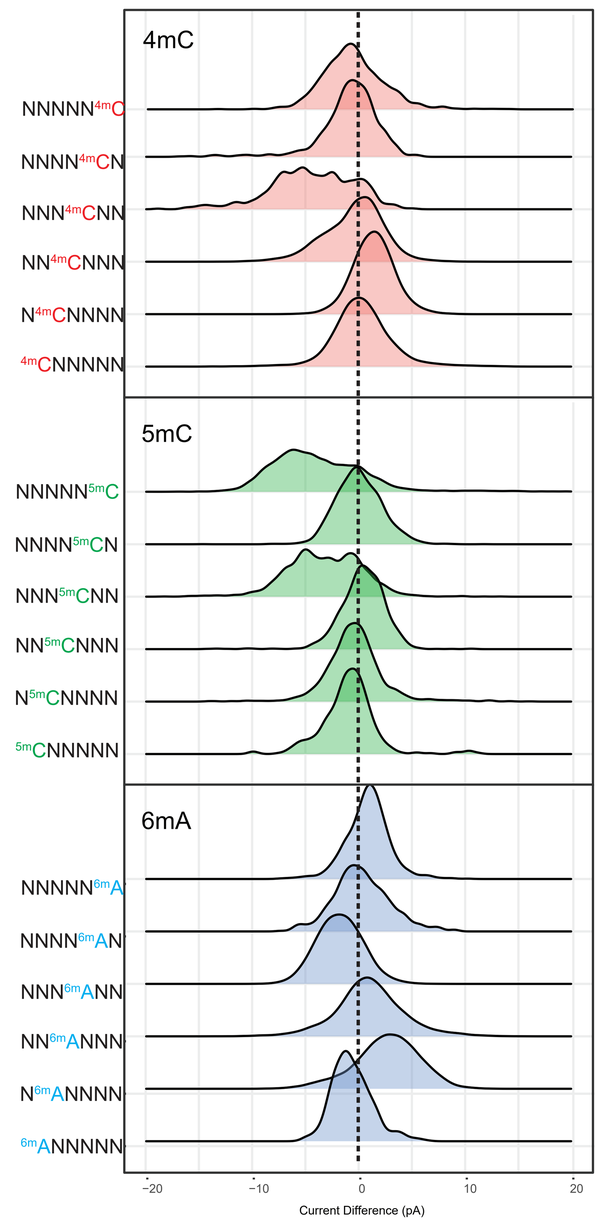


To perform methylation calling, we first align the basecalled data to a reference, then use eventalign to determine the electrical data which corresponds to each k-mer. From there, we can train our hidden markov model using the methylated and unmethylated samples to determine new model parameters, then realign our data. We typically perform 4 iterations of this training to determine the emission states for k-mers containing methylated bases.

Results

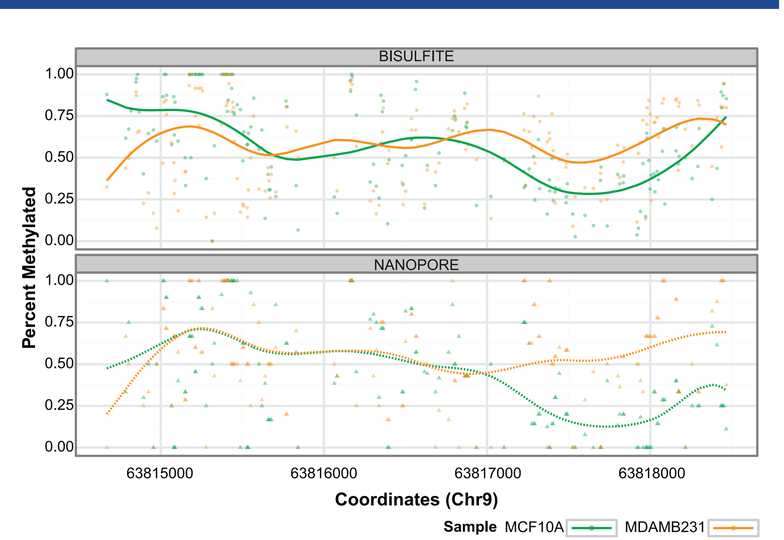


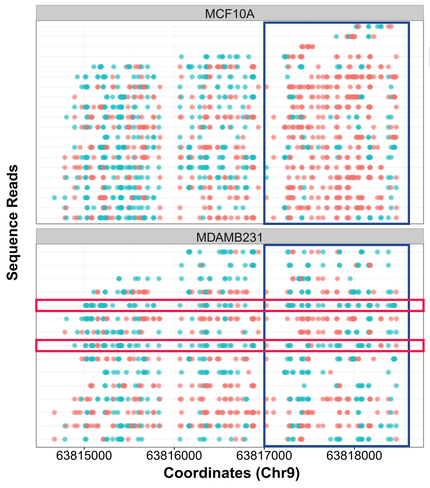
Base modifications can shift the current signals from the unmethylated signal depending on both context and modification type. Pictured above are current distributions trained for several different k-mers using methyltransferases M.HinfI (G6mANTC), dam (G6mATC), or Sin395ORF667 (GAT5mC). Though the dam and sin395 enzymes methylate the same motif (GATC), we can often distinguish the dam (6mA) methylation from the sin395 methylation (5mC).



MTase	Mod	Motif
PspJDRI	4-mC	(m4C)CGG
Sin395ORF667	5-mC	GAT(m5C)
Fnu4H	5-mC	G(5mC)NGC
M.SdeAII	5-mC	CCNGG(m5C)
M.HinfI	6-mA	G(m6A)NTC
BstXII (dam style)	6-mA	G(m6A)TC

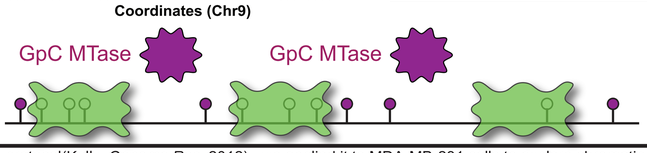
Using the enzymes summarized on the left, we trained and profiled how the position of methylation in a 6-mer affected the current profile. This is summarized above as a density plot of the differences in current between unmethylated and methylated k-mers. Interestingly, different modifications seem to have different locations of primary significance: 5-methylcytosine seems to show strong differences in many contexts when at the 6th position, while N6-methyladenosine shows signal at the 2nd and 4th position.



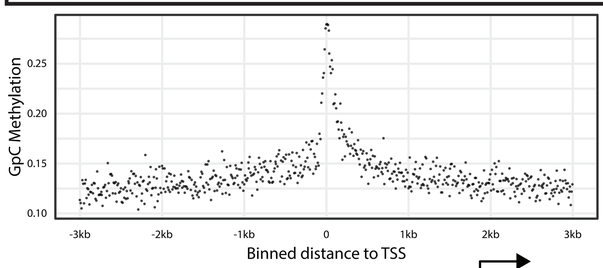


Using our existing 5-methylcytosine CG model, we can identify methylation in cancer versus normal samples. To leverage the relatively limited nanopore sequencing yield, we performed a digestion with MseI, followed by size selection for fragments between 3500-6000 bp. The resulting fraction of the genome (12 Mb) was then sequenced using both Illumina bisulfite sequencing and nanopore sequencing for MCF10A (breast epithelial normal cell line) and MDA-MB-231 (breast aggressive cancer cell line).

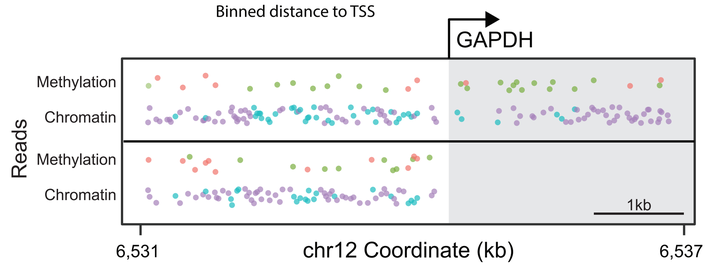
A selected CpG island is shown above, showing a differentially methylated region (DMR) on the right detected in both nanopore and illumina data. Importantly, on the left we can visualize individual reads, allowing us to see methylation patterns over a ~2kb region. Specifically, note the highlighted (red) reads - it seems that the hypermethylation in cancer samples occurs over a large range, though only resulting in a local DMR.

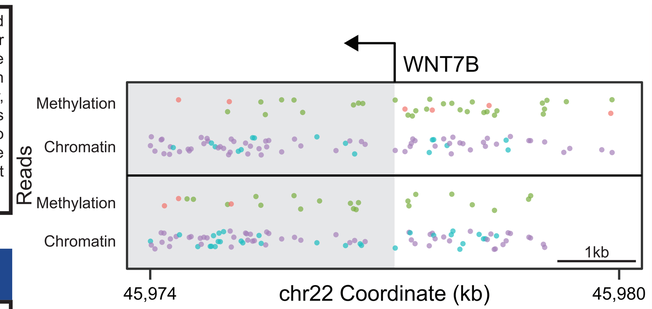


Adapting the NOMe-seq protocol(Kelly, *Genome Res.* 2012) , we applied it to MDA-MB-231 cells to explore chromatin and methylation state simultaneously using long-read nanopore sequencing. Specifically, we subjected cell nuclei to M. CviPI GpC methyltransferase treatment, which methylates GpC dinucleotides in accessible chromatin, while leaving inaccessible chromatin unmodified. The genomic frequency of GpC dinucleotides is high, providing sufficient resolution to resolve individual nucleosome positioning - but coupling it to nanopore sequencing provides this data along a single long (~10kb) molecule. We are already exploring using some of the other methyltransferases to multiplex this type of exogenous DNA labeling.



We performed an initial sequencing run of only one flow cell (1-2X coverage) to determine the effectiveness of the method and refine our analysis pipeline. First, we plotted the chromatin state as revealed by GpC methylation binned versus distance to the transcriptional start site (TSS). The profile fits accepted behavior, with chromatin generally most open near the TSS, then decaying as the profile moves away.





We examined individual read data around two genes - GAPDH, a constitutively expressed housekeeping gene, and WNT7b, a gene expressed in MDA-MB-231 cells. From the two reads we had covering the TSS of each gene, we called and plotted both natural CpG methylation (red/green) and exogenously applied GpC methylation (blue/purple). Gene bodies are shaded grey in the plot (left). Note the patterns of accessible and inaccessible chromatin upstream of the GAPDH TSS, potentially representative of nucleosomes, along with the good correlation of chromatin (GpC) and methylation (CpG). Inaccessible chromatin may also be due to TF binding or polymerase, so further work is still needed.

Acknowledgments

Thanks to NEB (Alexey Fomenkov and Rich Roberts) for sharing of bacteria and methyltransferases. We gratefully acknowledge conversations with Ben Langmead, Kasper Hansen, Nick Loman, Peter Thielen and Mark Akeson which contributed to this work and our understanding. We are also grateful to Oxford Nanopore for extensive technical support and information. Funding for this project provided by NHGRI (1R01HG009190). Disclosure: Winston Timp has two patents licensed to Oxford Nanopore