Nanodisco: Discovering and exploiting multiple types of DNA methylation from individual bacteria and microbiome using nanopore sequencing

Alan Tourancheau[1], Edward A. Mead[1], Xue-Song Zhang[2] and Gang Fang[1]

[1] Department of Genetics and Genomic Sciences and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[2] Center for Advanced Biotechnology and Medicine, Rutgers University-New Brunswick, Piscataway, NJ, USA

Abstract

Nanopore sequencing provides a great opportunity for direct detection of chemical DNA modification. However, existing computational methods were either trained for detecting a specific form of DNA modification from one, or a few, specific sequence contexts (*e.g.* 5-methylcytosine from CpG dinucleotides) or for allowing *de novo* detection without effectively differentiating between different forms of DNA modifications. As a result, none of these methods supports *de novo,* systematic study of unknown bacterial methylomes. In this work, by examining three types of DNA methylation in a large diversity of sequence contexts, we observed that nanopore sequencing signal displays complex heterogeneity across methylation events of the same type. To capture this complexity and enable nanopore sequencing for broadly applicable methylation discovery, we generated a training dataset from an assortment of seven bacterial species and developed a novel method that couples the identification and fine mapping of the three forms of DNA methylation into a multi-label classification design. We evaluated the method and then applied it to individual bacteria and mouse gut microbiome for reliable methylation discovery. In addition, we demonstrated in the microbiome analysis the use of DNA methylation for binning metagenomic contigs, associating mobile genetic elements with their host genomes, and for the first time, identifying misassembled metagenomic contigs. This novel method has broad utility for discovering different forms of DNA methylation from bacteria, assisting functional studies of epigenetic regulation in bacteria, and exploiting bacterial epigenomes for more effective metagenomic analyses. Those methods are available through our new tool, **nanodisco**.

Preprint: https://www.biorxiv.org/content/10.1101/2020.02.18.954636v1
Nanodisco software, documentation, and tool showcase: https://github.com/fanglab/nanodisco

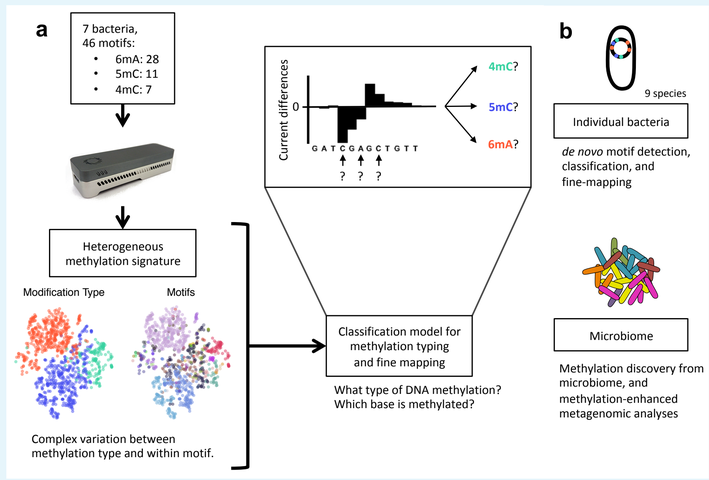

Figure 1: Schematics for method design and applications. (**a**) Using isolated bacteria with a wide variety of methylation motifs we explore the signal of DNA methylation in nanopore sequencing and characterize the major types of DNA methylation (4mC, 5mC, and 6mA). We observed a large variation and complex heterogeneity of current differences (native versus WGA) between methylation sequence context, which motivated us to develop a broadly applicable method for classifying DNA methylation into specific methylation type (4mC, 5mC, and 6mA) and fine mapping of the methylated base. (**b**) We performed comprehensive method evaluation and demonstrated the application of our method for methylation discovery from individual bacterial species (7 + 2 species) and microbiomes (methylation motif detection, classification, and fine mapping), as well as methylation-assisted metagenomic analysis (methylation binning and misassembly identification).

1. Heterogeneous signal variation induced by DNA methylation in nanopore sequencing

Generating methylation motif signatures: In order to comprehensively examine the variation of different types of DNA methylation within a broad scope of sequence context as measured by nanopore sequencing, we collected 46 well-characterized unique methylation motifs from a set of seven bacterial species with diverse methylation motifs (**Table 1**; see preprint for details). After sequencing both native and unmethylated (obtained after whole genome amplification, WGA) libraries, read events and associated current levels (picoampere, pA) were aligned to reference genomes using Nanopolish[1]. After normalization and filtering, current differences between native and WGA datasets were computed for each genomic position. To examine the variation of current differences across different DNA methylation types and motifs, we extracted current differences around each methylated base ([-6 bp, +7 bp]) and computed the methylation motif signatures (i.e. distribution of current differences at relative positions from the methylated bases, see **Fig. 2a**).

Organism name	Library type	Yield (bases)	Coverage
Bacillus amyloliquefaciens H	Native	792,809,181	186
	WGA	503,886,527	117
Bacillus fusiformis 1226	Native	806,872,641	152
	WGA	540,826,143	100
Clostridium perfringens ATCC 13124	Native	869,572,490	245
	WGA	598,911,051	128
Escherichia coli K-12 substr. MG1655	Native	989,844,145	200
	WGA	1,003,753,854	200
Helicobacter pylori JP26	Native	339,966,316	200
	WGA	337,927,240	200
Methanospirillum hungatei JF-1	Native	868,956,831	230
	WGA	428,552,409	111
Neisseria gonorrhoeae FA 1090	Native	445,185,393	194
	WGA	397,951,671	168

Table 1: Nanopore sequencing dataset coverage used for motif detection and classification. Native and WGA (from Whole Genome Amplification) libraries were prepared and sequenced on the MinION. Average coverages were computed using bedtools (version 2.26.0, parameters genomecov -d). *E. coli* and *H. pylori* datasets were downsampled to 200x.

Characteristics of methylation signal: The methylation motif signatures were further processed with t-SNE to visualize signal characteristics (**Fig. 2b**). A general clustering pattern is seen where methylation motif occurrences from the same methylation type tend to cluster together (**Fig. 2c**), although there are apparent overlaps. Importantly, we observed that current differences associated with different methylation motifs of the same methylation type often form different clusters, some individual motifs form distinct sub-clusters, *i.e.* current differences generally varies between different motifs of the same methylation type (e.g. T4mCTTC; **Fig. 2b,c**), and even between methylation events within the same methylation motif (e.g. GGW5mCC; **Fig. 2a,b**).

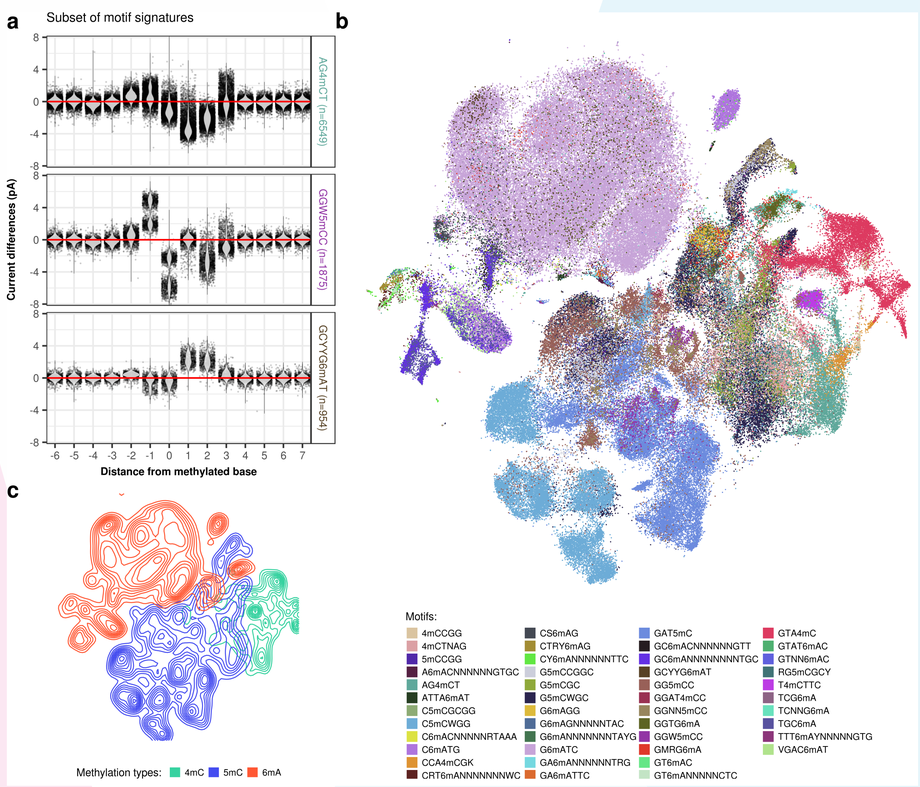

Figure 2: Systematic examination of three main types of DNA methylation with nanopore sequencing. (**a**) Variation of current differences across methylation occurrences as illustrated by motif signatures from three motifs (AG4mCT, GGW5mCC, and GCYYG6mAT). For each motif, current differences near methylated bases ([- 6 bp, + 7 bp]) from all isolated occurrences are plotted with conservation of relative distances to methylated bases. Distributions of current differences for each relative distance are displayed as a violin plot. Current differences axis is limited to -8 to 8 pA range. (**b**) Variation of current differences across methylation occurrences as illustrated by projection with t-SNE for 46 well-characterized motifs. Each dot represents one isolated motif occurrence colored by methylation motif. For each motif occurrence, current differences from 22 positions near methylated bases ([- 10 bp, + 11 bp]) were used. (**c**) Similar to **b** but colored by DNA methylation type with additional processing to reveal cluster density indicated by relief.

Further analysis of motif signatures suggests that this across-motif and within-motif variation can be in a large part explained by sequence variation from degenerated position in motifs as well as sequences flanking the consensus motifs (**Fig. 3**). **Fig. 3a** shows an example where signature sub-clusters for a 5mC motif (GGW5mCC) can be partially explained by sequence diversity near methylated bases (within-motif sequence variation). Similar observations were made with respect to sequence variation outside of consensus methylation motif (GAT5mC; **Fig. 3b**).

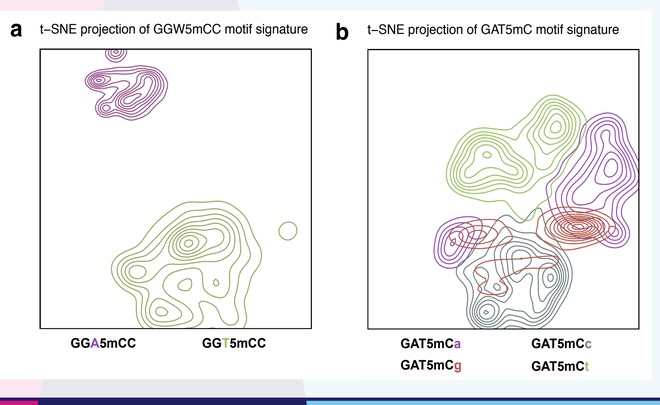

Figure 3: Local sequence context effect on motif signatures. (**a**) independent t-SNE projection of GGW5mCC motif occurrences with cluster density displayed as relief. Clusters are colored according to degenerated base within the methylation motif. (**b**) Another example of sequence-dependent variation for GAT5mC motif occurrences displayed after independent t-SNE projection with cluster density displayed as relief. Clusters are colored according to the first base following GAT5mC motif.

2. A new method that enables de novo methylation typing and fine mapping

Methylation motif detection: We build on existing methods[1, 2, 3]. In brief, 1) current levels are compared between native and WGA datasets for each genomic position; 2) p-values are combined locally with a sliding window-based approach followed by peak detection; 3) flanking sequences around the center of peaks are used as input for MEME motif discovery analysis. Overall, 45 of the total 46 well-characterized methylation motifs from the seven bacteria were successfully re-discovered.

***De novo* methylation motif typing and fine mapping** (see preprint for details): Methylation type classification (i.e. identify the type of DNA methylation) and methylation fine mapping (i.e. identify the position of the methylated base) are coupled problems that need to be approached simultaneously. Although the methylated base is not always at the center of the current differences (**Fig. 2a**), we did observe a relatively narrow window of no more than +/- 3 bp offsets from peak centers across the 46 well-characterized motifs. This motivated us to design a novel multi-label classifier training strategy in which each well-characterized methylation occurrence is represented by multiple features vectors (of length 12) with offsets relative to the known methylation position (+/- 3bp; **Fig. 4a**). Each methylation occurrence from a wide range of sequence context is learned 7 times by the classifier, each time using current differences at a specific offset from the methylated base (**Fig. 4a,b**). For a given test sample with unknown methylation type and unknown methylated position, the classifier will first use the center of current differences as an approximation of the methylated position and then predict the methylation type and the exact methylated position. This is the core design of our method that enables completely *de novo* methylation typing and fine mapping, which is critical for practical applications to unknown bacterial genomes. Performances from nine different classifiers were evaluated using leave-one-out cross validation strategy (LOOCV; **Fig. 4c**). Overall, k-nearest neighbors, random forest, and neural network had relatively better performances with at least 95.7% of motifs correctly typed and fine mapped (**Fig. 4d**).

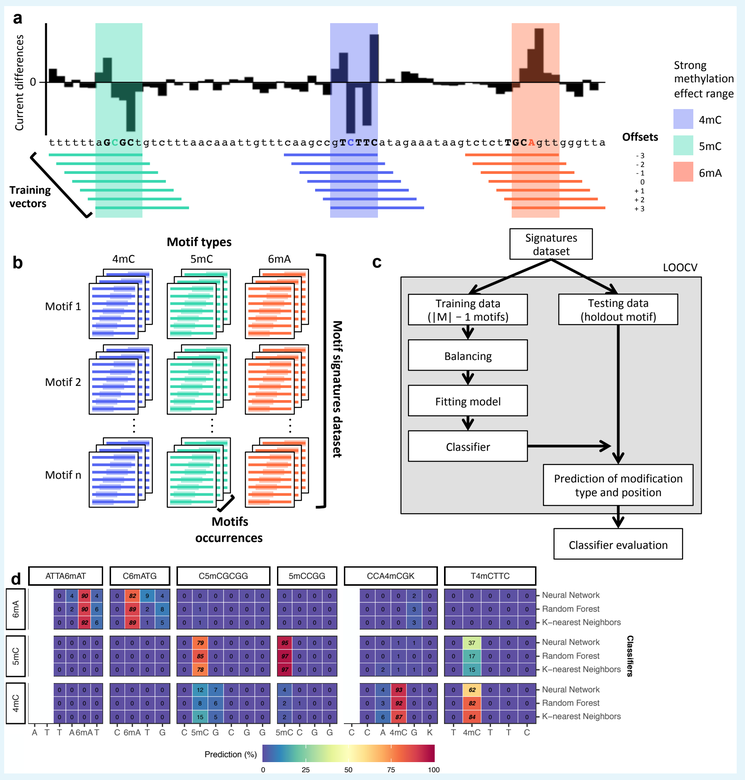

Figure 4: Classification and fine mapping of three types of DNA methylation. (**a**) Schematic representation of dataset building for classifier training. (**b**) Each training vector is labeled with the corresponding methylation type and offset used (current differences flanking 183,818 methylated bases from 46 distinct motifs were used). (**c**) Classifiers performances were evaluated using LOOCV. (**d**) Subset of LOOCV classifier evaluation results from 6 motifs.

3. Methylation discovery from microbiome and methylation-enhanced metagenomic analyses

Methylation binning (see preprint for details): Metagenomic assembly often results in fragmented genomes where contigs are short hence including only a limited number of occurrences of each motif. This makes methylation motifs discovery statistically underpowered if each metagenomic contig is examined separately. Recent work by Beaulaurier *et al.* demonstrates that microbial 6mA DNA methylation can be exploited to enhance the grouping of metagenome contigs (i.e. methylation binning) using SMRT sequencing[4]. Instead of trying to discover precise methylation motifs from individual contigs, methylation features are computed for each contig and those are grouped into bins based on methylation profiles similarities. We developed a new methylation binning approach specifically for nanopore sequencing data considering the fundamental differences from SMRT sequencing: 1) the methylation signal span multiple events near methylated bases (**Fig. 2a**) rather than confined to a single base as in SMRT sequencing, 2) the sensitivity to 5mC in addition to 6mA and 4mC (weak 5mC signal in SMRT sequencing). To summarize, after *de novo* assembly of the metagenome using nanopore sequencing data only, current differences are computed and further processed to construct a methylation profile matrix (**Fig. 5a**; see preprint for details). The methylation profile matrix contains methylation features from informational motifs, which can then be used in a clustering analysis based on the similarity of methylation profile among contigs.

We apply our methylation binning method to MGM1, a mouse gut microbiome sample (**Table 2**). The initial automated binning revealed ten bins which were further refined by three round of per-bin motif detection followed by guided methylation binning (i.e. using *de novo* discovered methylation motifs). The final methylation binning round was performed using 80 *de novo* detected methylation motifs and revealed thirteen bins containing from 3 to 43 contigs in each (**Fig. 5b**).

Sample name	Library type	Yield (bases)	Number of contigs	Contigs N50	Longest contig
MGM1	Native	5,022,756,117	2,905	47,250	1,116,461
	WGA	3,140,343,539			

Table 2: Nanopore sequencing dataset for microbiome analysis. MGM1 is mouse gut microbiome sample.

Mobile genetic element binning: Methylation binning also permit to link mobile genetic elements (MGEs) to their host genome as they share methylated motifs and therefor methylation profiles. We were able to bin 4 of the 8 annotated MGEs from MGM1 sample according to their methylation profiles (**Fig. 5b;** 11 out of 19 from SMRT assembly not shown).

Detection of misassembly (see preprint for details): In a nutshell, the methylation pattern is expected to be largely consistent across different regions of an authentic metagenomic contig. Following this rationale, we computed scores representative of the methylation status for *de novo* detected motif along the contigs. We discovered two contigs from SMRT sequencing based metagenomic assembly of the MGM1 sample showing inconsistent intra-contig methylation status (**Fig. 6,** 2nd contig not shown). By comparing methylation patterns from methylation motif sets from the other bins, we found that the contigs in question are chimeric contigs from two distinct Bacteroidales species.

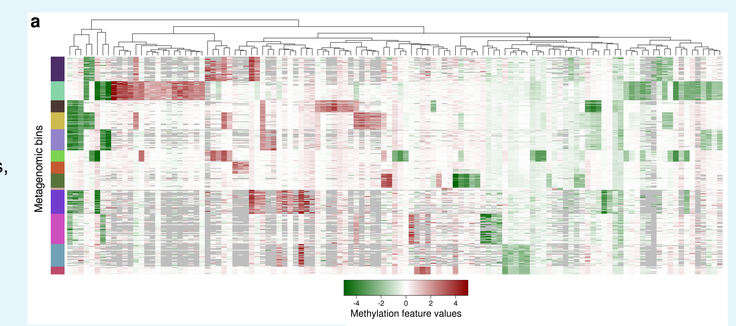

Figure 5: Methylation analysis of mouse gut microbiome samples. (**a**) Heatmap of the methylation profile matrix using a subset of methylation feature values computed across binned contigs from MGM1 sample. Significant methylation features (|pA| > 1.5) were computed from the 80 *de novo* detected motifs in the 13 bins (see **b**). Missing methylation features from contigs (less than 5 motif occurrences) are colored in grey. (**b**) Methylation binning of MGM1 metagenome contigs using *de novo* discovered motifs (after three rounds of binning followed by motif discovery). Methylation features computed from *de novo* discovered motifs are projected on two dimensions using t-SNE. Contigs are colored based on bin identities with point sizes matching contig length according to the legend.

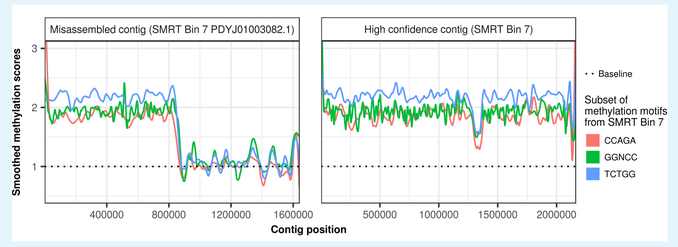

Figure 6: Misassembly detection using methylation motif information along contigs. Left panels: misassembled SMRT Bin 7 contig (PDYJ01003082.1). Right panel: properly assembled SMRT Bin 7 contig (PDYJ01000763.1). Using a subset of *de novo* detected motifs from Bin 7, a score is computed for each motif occurrence by taking the average of absolute current differences from six consecutives positions with the most perturbation. Smoothed methylation scores are consistent in the high confidence contig (right panel), while a switch of methylome occurs near 800 kbp supporting the existence of misassembly.

Conclusion

- DNA methylation of the same type have great variation and heterogeneity in nanopore sequencing suggesting that a broadly applicable method for methylation discovery is best trained using a comprehensive dataset with methylation motif diversity rather than a dataset of one or few specific motifs.
- **Nanodisco** enables the *de novo* characterization of unknown bacterial methylomes from both individual bacteria and microbiome samples (95.7% of motifs correctly typed and fine mapped) with our new methylation motif classification-based method.
- **Nanodisco** enables methylation binning of metagenomic contigs and linking of MGEs to host genomes building on the method reported for SMRT sequencing.
- **Nanodisco** enables the identification of misassembly in metagenome by visualizing methylation patterns along assembled metagenomic contigs.

Acknowledgments

We thank Alexey Fomenkov and Sir Richard J. Roberts for their help with the bacterial strain selection and for providing us with DNA samples (*B. amyloliquefaciens*, *B. fusiformis*, and *N. otitidiscaviarum*). We also thank Robert Gunsalus (*M. hungatei*), Susan Logan (*C. perfringens*), Lydgia Jackson (*N. gonorrhoeae*), Bernhard Schink, Nicolai Müller, and Anja Keller (*T. phaeum*) for providing us with DNA samples. We thank Yimeng Kong and Mi Ni for providing helpful feedback for early versions of the manuscript. The work was supported by a seed fund from Icahn Institute for Genomics and Multiscale Biology (G.F.), and by R01 GM128955 (G.F.) from the National Institutes of Health. G.F. is a Hirschl Research Scholar by Irma T. Hirschl/Monique Weill-Caulier Trust, and a Nash Family Research Scholar. This work was also supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

References

1. Simpson, J.T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* 14, 407-410 (2017)
2. Stoiber, M. et al. De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv* (2017)
3. Bailey, T.L., Johnson, J., Grant, C.E. & Noble, W.S. The MEME Suite. *Nucleic Acids Res* 43, W39-49 (2015)
4. Beaulaurier, J. et al. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol* 36, 61-69 (2018).

Full list of references available in our bioRxiv preprint: 10.1101/2020.02.18.954636v1