Supplement Information

Nature of nanopore sequencing signal from Oxford Nanopore Technologies

Raw nanopore signal corresponds to electric current level (pA) sampled at 4000 hz across the nanopore while a DNA strand is transferred from one compartment to the other in a 450 bp.s-1 ratcheting motion. Higher order of signal structure, called events, consists in consecutive signal level corresponding to multiple measures of current for a specific relative position of the DNA strand inside the pore. The initial signal processing performed by the base caller, Albacore (version 2.3.4), detects those consecutive events and translates them into a nucleotide sequence.

Additional information for methylation motif validation

Our *de novo* methylation motif detection analysis also discovered six motifs not included in the confident list. Two motifs were detected in *H. pylori* (*i.e.* GGWTAA and GGWCNA, likely 6mA on sixth position) but the analysis of SMRT sequencing data suggest that they are partially methylated. Two additional motifs were found in *N. gonorrhoeae.* One of them is GTANNNNNCCC, likely modified by the MTase of GT6mANNNNNCTC, but SMRT data shows weak methylation signal suggesting that the motif is partially methylated (not methylated in all reads). The other one is TCACC, a 5mC methylation motif according to our classification (*i.e.* T5mCACC), which would explains why it was not detected with SMRT sequencing analysis. Finally, YGGCCR and WGGCCW were discovered in *B. fusiformis* and *C. perfringens* respectively. While both were expected to be the non-degenerated methylation motifs GG4mCC, SMRT sequencing data analysis also suggests that the others subsets of motifs (non-YGGCCR or non-WGGCCW) were not fully methylated explaining our results.

Other unconfident methylation motifs were found only with SMRT sequencing. In *H. pylori*, we listed three unconfident motifs (*i.e.* CTGG6mAG, CCTCT6mAG, and STA6mATTC) with weak signals suggesting that they were false discovery or partially methylated motifs (not methylated in all reads), thus not suitable for training. However, we also found a methylation motif in *N. gonorrhoeae* with strong SMRT sequencing signal (*i.e.* CC6mACC) while little to no sign of methylation are visible with ONT analysis (*i.e.* no perturbation in average current differences near motif). It's unclear if this particular methylation motif is not detected because ONT method is not sensitive to change in nucleotide (between A and 6mA) in CCACC sequence context or because it's not methylated in our *N. gonorrhoeae* sample thus it was not used in our analysis.

Note that all the ambiguous motifs mentioned in this section were treated as potential methylation motifs when removing overlapping signal in order to avoid possible compound effects. However, they were ignored from all analysis.

Evaluation of different signal processing methods for motif detection and characterization

We explore other potential sources of variation by comparing general signal characteristics using t-SNE projection with different base callers, and different signal processing workflows. The base caller versions tested (Albacore 1.1.0 vs Albacore 2.3.4 vs Guppy 3.2.4) did not significantly change the signal characteristics (**Extended Data Fig. 3a**) and give consistent methylation detection performances (a slight improvement of 0.1% from Albacore 1.1.0 to Guppy 3.2.4 in term of the area under the Precision-Recall curve; **Extended Data Fig. 3b**). Similar methylation signal properties were obtained from bacterial datasets processed with Tombo compared to our pipeline, suggesting that no significant bias was introduced: motif signatures are similar (**Fig. 2a and Supplementary Fig. 1a**), methylation sites cluster by type (**Fig. 2c** and

Supplementary Fig. 1c), by motif (**Fig. 2b and Supplementary Fig. 1b**), and by local sequence context (**Fig. 3b and Supplementary Fig. 1d**). The outlier removal procedure reduces noise in the motif signatures (**Extended Data Fig. 3c-e**) and does not introduce significant bias as indicated by the high similarity with signal processed with Tombo (**Fig. 2a** and **Supplementary Fig. 1a**). Finally, we also confirmed that current differences obtained using a *de novo* assembled genome (*E. coli* at 200x; Methods) were consistent with the one obtained from matching reference genome ruling out the possibility that observed clustering pattern could be explained by an inaccurate reference genome (**Extended Data Fig. 3l,m**).

In addition, we observed similar method performances when we used a *de novo* assembled genome (*E. coli* sample at 200x, assembled with 99.94% genomic consensus accuracy) for motif detection (**Extended Data Fig. 3m**), and for motif typing and fine mapping (**Extended Data Fig. 3n**). We also evaluated the impact of genomic coverage using subsampled datasets of *H. pylori* and observed improvement in motif enrichment (from 5x to 75x and then plateaued; **Extended Data Fig. 7a,b**), as well as in motif typing and fine mapping across studied motifs (from 10x to 30x; **Extended Data Fig. 7c**). Furthermore, we evaluated the impact of genomic motif frequency (i.e. number of motif occurrences per Mb of genome sequence) on motif enrichment performance by creating *in silico* datasets with a wide range of motif frequencies (**Extended Data Fig. 7f**).

Limiting factor for methylation motif detection

Genomic coverage strongly affects methylation motif detection ability with substantial improvement in motifs enrichment up to 75x in *H. pylori* with 20% to 90% of motif detected by increasing coverage from 5x to 75x (**Extended Data Fig. 7a,b**). Overall, 75x

(37.5x per strand) is sufficient to detect 100% and 90% of motifs in *E. coli* and *H. pylori* respectively. In addition, we observed variation in enrichment across motifs even when variation in motifs frequency was accounted for (**Extended Data Fig. 7d,e**). Motif specific performances depend on the amount of current perturbation introduced by the methylation compared to the non-methylated signal. For example, the G6mAGG motif signature displayed weak current differences and was not detected for *H. pylori* dataset at lower coverage (<20x). At lower coverage, undetected motifs can display a clear signature although not sufficient to be enriched enough to detect them. Finally, in practice, bacterial methylation motifs have various frequencies in genomes sometimes independent of their complexity, which seems to be a limiting factor for their detection (e.g. GT6mAC in *H. pylori*, **Extended Data Fig. 7f**). Note that while methylation motif signatures represent how DNA methylation affect ionic current in a specific genomic context during sequencing, some of their characteristics depend on the data processing method used (*e.g.* base caller, reads mapper, event aligner, and normalization). We expect that methylation motif detection performance will increase with improvement of nanopore sequencing preprocessing methods, notably for base calling and signal alignment to a reference sequence.

Approximation of methylated position from motif signature

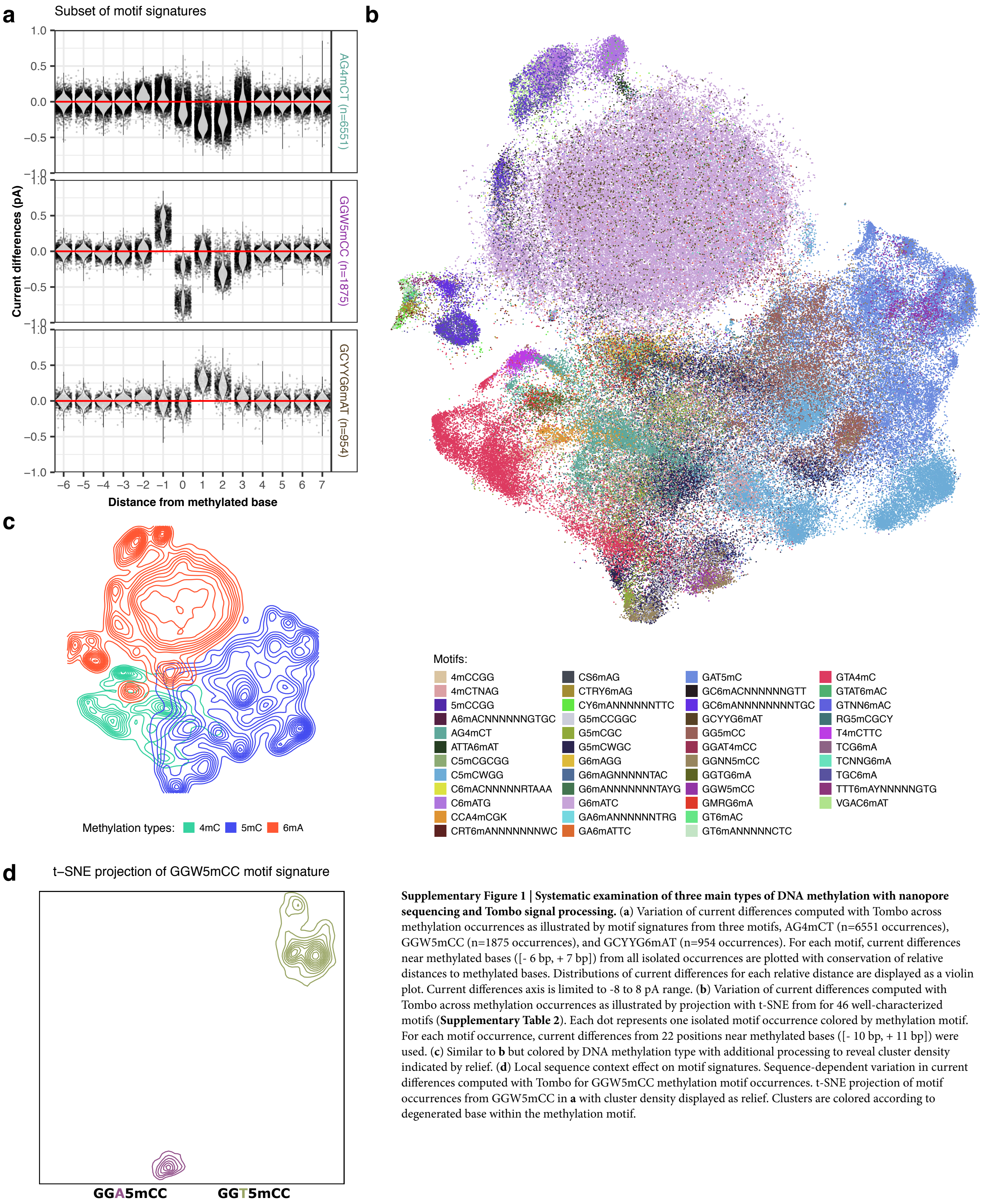
Our current method for approximating methylated position within *de novo* detected motifs relies on the identification of the center of the motif signature. However, other educated guesses could be made based on motif signature and refining plots, which would permit reducing the DNA methylation position research space. First, main current differences are in the [- 2 bp, + 3 bp] range from the methylated base (**Extended Data Fig. 1**) meaning that for bipartite motifs one could ignore part of the motif depending on which specificity subunit is aligned with current differences. Similarly, this could be done for long motifs if current differences are at one of the motif extremities. This phenomenon is indirectly used in our approximation approach. Second, motif signatures display important variation when the methylated base is close to non-fixed bases, *i.e.* next to a degenerated base or near motif extremities. This strategy was not used in the current implementation.

Mock microbiome from individual bacteria

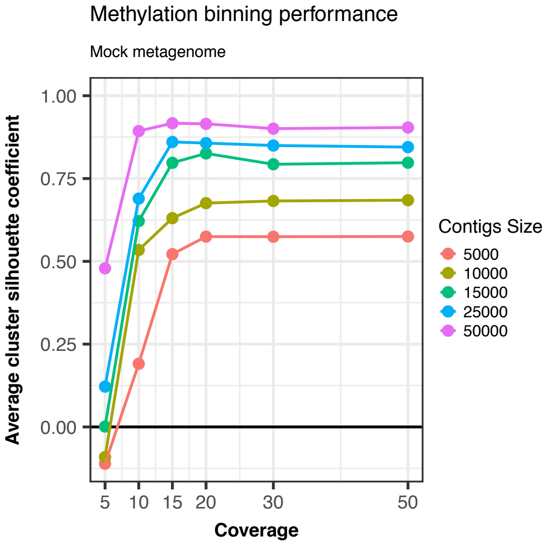
In order to define motif selection procedure for contig methylation binning, we constructed a mock metagenome assembly from our individual bacteria reference genomes (n=7). Reference genomes were fragmented following mouse gut metagenome contig length distribution from previous SMRT study[1]. Nanopore sequencing native and WGA datasets subsampled at a coverage of 50x were then mapped on the mock metagenome assembly and processed similarly to individual genomes to generate current differences and associated U test p-values (Methods). Possible methylation motifs from the initial set (n=210,176) are scored for long contigs (>=500 kbp) according to the procedure described in Methods. Rules for methylation motif features selection were defined to enrich the final list in known methylation motifs from bacteria in the mock community. Only genomic positions with 10x coverage were used in both scoring steps.

We applied the following cutoff on methylation features: minimum absolute current differences (1.5 pA), minimum number of motif feature occurrences per confident contigs (20), minimum number of significant features in bipartite motifs (2), and discard overlapping motifs (bipartite motif explained by 4 to 6-mers motifs). Any motif features satisfying those requirements are scored in remaining contigs. Missing methylation features and those computed from fewer than 5 motifs occurrences were set to small random pseudovalues in the [-0.2, 0.2] range to reduce correlation from missing methylation features.

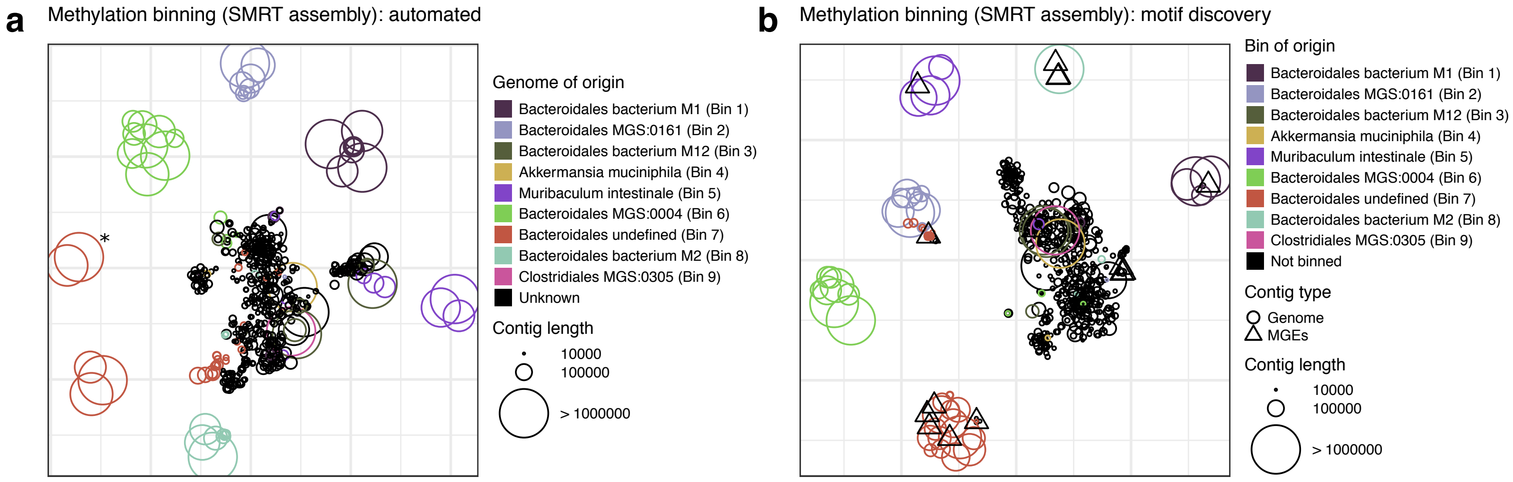
We further evaluated the impact of coverage on methylation binning performance by generating a mock metagenome from individual datasets subsampled at 6 depths (genomic coverage of 5x, 10x, 15x, 20x, 30x, and 50x). Similarly, we also evaluated the impact of contig length on binning by creating a mock metagenome with contigs of length varying from 5 kbp to 50 kbp. For each combination of the contig lengths and sequencing depths, we computed methylation features for the expected methylation motifs. To control for the motif frequency variation across contigs of the mock metagenome, we filtered out inconsistent features (i.e. features with absolute value > 0.5 pA, which are present in more than 15% but less than 85% of the contigs) effectively eliminating the confounding factor. Finally, we performed t-SNE based dimensionality reduction followed by quantitative evaluation using average cluster silhouette coefficients[2]. Cluster silhouette coefficients were computed for each species (n=7) at each contig length and depth combination. A coefficient of -1 indicates complete mixing of contigs between species, while a cluster silhouette coefficient of 1 indicates complete separation. We observed a clear improvement of average cluster silhouette coefficients at 10x compared to 5x for all contig length considered (**Supplementary Fig. 2**). Furthermore, as expected, as contig length increases, the performance of methylation binning also increases because more motif instances make the methylation feature estimation more accurate, reducing noise and making common methylation profiles easier to group together.



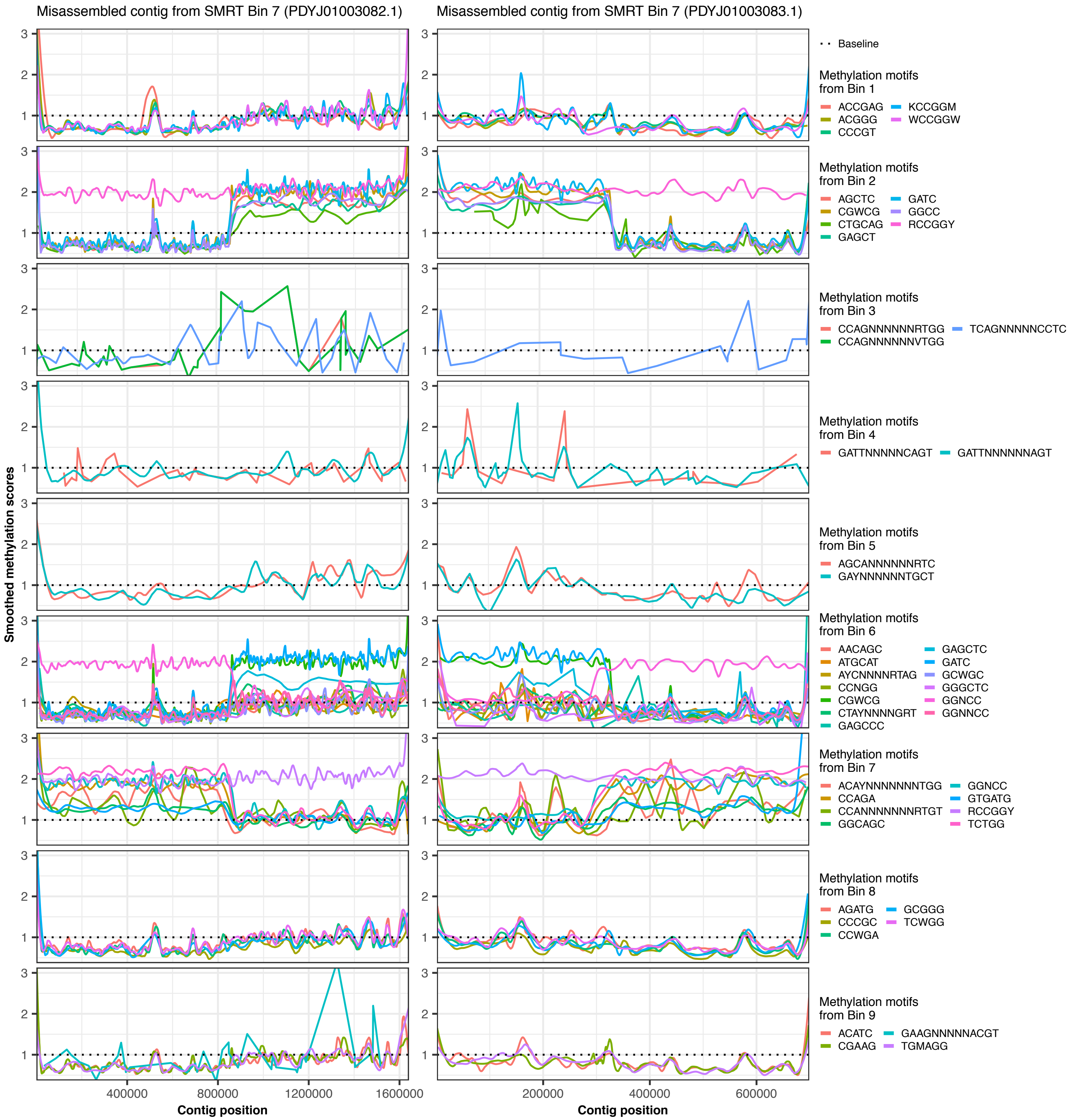
Supplementary Figure 1 | Systematic examination of three main types of DNA methylation with nanopore sequencing and Tombo signal processing. (**a**) Variation of current differences computed with Tombo across methylation occurrences as illustrated by motif signatures from three motifs, AG4mCT (n=6551 occurrences), GGW5mCC (n=1875 occurrences), and GCYYG6mAT (n=954 occurrences). 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 computed with Tombo across methylation occurrences as illustrated by projection with t-SNE from for 46 well-characterized motifs (**Supplementary Table 2**). 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. (**d**) Local sequence context effect on motif signatures. Sequence-dependent variation in current differences computed with Tombo for GGW5mCC methylation motif occurrences. t-SNE projection of motif occurrences from GGW5mCC in **a** with cluster density displayed as relief. Clusters are colored according to degenerated base within the methylation motif.



Supplementary Figure 2 | Theoretical performance of methylation binning on mock microbiomes. Impact of contig coverage and contig length was assessed by generating mock metagenomes from individual bacteria datasets subsampled at 6 depths (genomic coverage of 5x, 10x, 15x, 20x, 30x, and 50x) with contig lengths varying from 5 kbp to 50 kbp (Supplementary text). Cluster silhouette coefficients were computed for each species (n=7) at each contig length and depth combination. A coefficient of -1 indicates complete mixing of contigs between species, while a cluster silhouette coefficient of 1 indicates complete separation.



Supplementary Figure 3 | Methylation analysis of MGM1 sample with SMRT metagenome assembly. (**a**) Automated methylation binning of MGM1 metagenome contigs (without precise methylation motif discovery). Methylation status of common motifs (n=210,176) was screened across large contigs (>=500 kb) through computation of methylation feature vector (Methods, **Extended Data Fig. 8**). Informative motifs were selected and their status evaluated across remaining contigs. Resulting methylation features are projected on two dimensions using t-SNE. Contigs are colored based on bin identities assigned in the SMRT study1 with point sizes matching contig length according to legend. Our binning identified two contigs originally identified as Bin 7 clustered separately from the main bin (contigs marked with an asterisk) suggesting that they have a different methylation profile than remaining Bin 7 contigs, which was not observed with SMRT analysis. Contigs marked with an asterisk are used as example for misassembly detection in **Fig. 5d**. (**b**) Methylation based association of MGEs to host genomes. Annotation of potential MGEs was obtained from the per-bin reassemblies from the SMRT study1. Genomic contigs are colored by bin of origin with point sizes matching their length. Some contigs are now binned within different clusters than their bin of origin likely because the original contigs were chimeric. Note that many contigs from Bin 7 reassemblies are affected because of the two major misassembled contigs identified.



Supplementary Figure 4 | Detection of misassemblies in Bin 7 contigs from methylation motif signal. Identification of contamination origin for the two contigs mislabeled as Bin 7 (PDYJ01003082.1 and PDYJ01003083.1, marked with an asterisk in **Supplementary Fig. 3a**). We scored occurrences from methylation motifs found in each bin separately and smoothed signal along misassembled contigs (Methods). Scores from motif occurrences overlapping Bin 7 motifs were removed. Scores from Bin 2 motifs are consistently high in the second half of contig PDYJ01003082.1 and first half of contig PDYJ01003083.1 suggesting contamination originated from Bin 2 genomic sequences.

Reference

1. Beaulaurier, J. et al. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol* **36**, 61-69 (2018).
2. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53-65 (1987).