Title: **Differential adenine methylation analysis reveals increased variability in 6mA in the absence of methyl-directed mismatch repair**

Authors: Carl J. Stone[1,2], Gwyneth F. Boyer[3], and Megan G Behringer[1,2,4*]

Affiliations:

[1]Department of Biological Sciences, Vanderbilt University, Nashville, TN

[2]Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN

[3]Biodesign Center for Mechanisms of Evolution, Arizona State University, Tempe, AZ

[4]Department of Pathology, Microbiology, and Immunology, Vanderbilt University, Nashville, TN

Keywords: bacterial methylation, N6-methyladenine, 6mA, epigenetics, genomics, DNA modification, experimental evolution, *Escherichia coli*

ABSTRACT

Methylated adenines (6mA) are an important epigenetic modification in bacteria, but challenges in accurate and high-throughput detection of 6mA have previously limited the ability to conduct comparative studies. Nanopore sequencing has allowed for accurate and cost-effective characterization of a variety of DNA modifications; however, most bioinformatics tools for Nanopore data are designed for CpG methylation, and there are currently no tools designed for comparative 6mA methylomics studies. In this work we present CoMMA, a pipeline for analyzing differential methylation of 6mA modification at a GATC motif. We applied the CoMMA pipeline to *Escherichia coli* clones that were experimentally evolved for 2400 generations to show how bioinformatic 6mA characterization is complementary to genetic sequencing and informative on its own. We show that, in the absence of methyl-directed mismatch repair, *E. coli* GATC sites are overall hypomethylated and display increased methylation variability, indicative of reduced selection and increased drift. The CoMMA pipeline is a reliable tool to be used with the growing number of Nanopore methylation callers.

TEASER

A novel bacterial differential methylation pipeline reveals that methyl-directed mismatch repair in *Escherichia coli* maintains 6mA genome-wide.

INTRODUCTION

DNA methylation can alter the expression and regulation of DNA without changing its sequence. The advent of third-generation sequencing has made the identification of DNA base modifications easier than ever before. Previous methods for characterizing methylated DNA bases suffered from two major shortcomings, as they were either low-throughput, like mass spectrometry- or microarray-based methods; or relied on PCR-based technologies and necessitated multiple sequencing runs that were specific to the chemistry of a particular DNA base modification, as in bisulfite sequencing for 5-methylcytosine (5mC) modification[1–3]. In contrast, PCR-free third-generation sequencing technologies, such as Nanopore sequencing (Oxford Nanopore Technologies, or ONT) and single-molecule real-time (SMRT) sequencing (PacBio), sequence single DNA molecules by detecting variation in an ionic current trace or fluorescent pulse, respectively, that can be linked to the nucleotide's chemical structure[3,4]. Using these technologies, we can now infer the presence of varied DNA base modifications in a high-throughput manner and with only the methylated base motif as prior knowledge. Nanopore sequencing is an accurate and accessible method for methylated base detection in bacteria; however, most bioinformatics tools have been developed for the analysis of 5mC, a relatively rare modification in proteobacteria that is focused on due to its relevance to human cancer[1,4–8]. While most methylation-calling programs are shown to be capable of detecting modified bases other than 5mC, relatively little work has been done assessing the accuracy of identifying these other bases in real biological samples, and there is a dearth of programs designed to compare modified bases within and between bacterial samples.

Bacterial epigenetics, primarily in the form of 6mA, has an outsized role in the regulation of various cellular processes[9–15]. The modified DNA bases found in bacteria and archaea include 6mA, 5mC, and N^4-methylcytosine (4mC)[9,12,16]. These modifications are added by DNA methyltransferases, either as part of a restriction-modification system or by orphan methyltransferases[17–19]. In gammaproteobacteria, 6mA modified by the orphan methyltransferase Dam at GATC motifs is the most abundant modified base in the genome[20–22]. There are 19,210 of these palindromic GATC motifs in the *Escherichia coli* K-12 genome, or one GATC site on average every ~250 bp (Fig S1)[23]. 6mA influences varied processes including, but not limited to, replication

timing, transposon mobility, gene expression, virulence, and methyl-directed mismatch repair (MMR)[9,14,24–35]. Many of these phenotypes occur as bistable states where stochasticity affects methylation status (methylated, unmethylated, or hemimethylated) at a regulatory locus, and proteins that preferentially bind based on methylation status maintain that status as a positive feedback loop[36]. One example of bistability under control of 6mA is the *pap* operon in uropathogenic *E. coli* which encodes adhesive fimbriae during infection[37,38]. Dam competes with global regulator Lrp at the *pap* promoter, and cells with methylated promoters exclude Lrp binding at the methylated site and express regulatory proteins that initiate a positive feedback loop by promoting Lrp binding at nearby sites that do not interfere with *pap* expression[37–39].

The most significant impact of 6mA on the *E. coli* genome is its necessity for the function of MMR. GATC sites are nearly all methylated on both strands, except for the brief period after chromosome replication before the daughter DNA strand is methylated by the orphan methyltransferase Dam[26,40]. When a mismatch is detected by MMR factors (MutHLS), the nearby hemimethylated GATC sites are used to identify the daughter strand and resolve the mismatch[26,30,41,42]. Inactivation of any of *mutHLS* leads to a ~150-fold increased mutation rate in *E. coli*[43]. Since MMR relies on the presence of hemimethylated GATC sites, alterations to the function of Dam in *E. coli* and other gammaproteobacteria also reduce the effectiveness of MMR[25,29,44].

Given the role of MMR in preventing replicative errors from becoming permanent mutations, we hypothesize that MMR imposes selective pressure on the *E. coli* genome to maintain GATC sites and associated 6mA modifications. While 6mA can have local gene expression effects when coincident with regulatory elements and transcription factors, MMR is likely the most significant use of (and influence on) 6mA genome-wide. Thus, we hypothesize that in the absence of MMR, *E. coli* experimentally evolved for thousands of generations will exhibit increased variability in per-site 6mA methylation and a genome-wide decrease in 6mA due to a sustained accumulation of epimutations that do not incur the strongly deleterious fitness effect of causing MMR dysfunction. There has been limited examination of the dynamics of 6mA in experimentally evolving bacterial populations[45,46]. Thus, in this study, we sought to investigate how the presence and absence of MMR influence 6mA methylation in evolving bacterial populations over an extended time period. To accomplish this, we created a reproducible 6mA differential methylation-calling pipeline:

CoMMA (Comparison of Microbial Methylated Adenines). As part of this pipeline, we first characterized genome-wide methylation frequency at GATC sites in *E. coli* MG1655. Then, focusing on clones isolated from evolved populations with initially WT and MMR-deficient genetic backgrounds, we applied CoMMA to characterize how 6mA methylation frequency changed among all 19,120 GATC motifs after 2000 generations of experimental evolution. We then compared our findings to studies of spontaneous mutation rates in *E. coli* to determine how MMR constrains genetic and epigenetic evolution at GATC sites.

RESULTS

Evaluation of methylation callers revealed large differences in precision.

Methylation-calling programs for Nanopore data have been thoroughly benchmarked on the classification of 5mC. While popular programs are capable of identifying 6mA, there are limited studies where Nanopore data has been used on 6mA, especially in bacteria. Thus, we identified the location of the 38,236 GATC sites in the E. coli K-12 substr. MG1655 genome (**Fig S1**) and compared the reliability and reproducibility of commonly used methylation callers in characterizing 6mA at these sites. We sequenced DNA from three equal aliquots of a single *E. coli* K-12 MG1655 (strain PMF2) overnight culture on an ONT MinION portable sequencer. Nanopore sequencing produced an average read depth of ~50 reads per GATC site, and there was no significant strand bias in coverage (**Fig. S2**). We then used four different software packages (DeepSignal, mCaller, Megalodon, and Tombo) to calculate the proportion of methylated reads to total reads, called the GATC sites[4,6,47–49]. All four of these packages differ in their approaches and dependencies (**Fig. 1**)[4,6,47–49].

Each methylation caller produced varying distributions of predicted methylation values across all GATC sites (**Fig. 2A)**. Tombo classified a high proportion of sites as completely methylated (all reads were classified as methylated at 31.8% of sites). This is consistent with previous benchmarking on 5mC methylation where Tombo over-predicted methylation at CpG sites[5]. DeepSignal and mCaller predicted fewer sites as methylated (median percent methylated reads of 87% and 83%, respectively). Megalodon identified a median percent methylation of 97% across all sites.

We next determined the precision of each methylation caller by comparing how consistently each GATC site was called as methylated across the three sequencing replicates. We considered a GATC site to be completely methylated if ≥60% of reads on both strands were inferred to be methylated. Alternatively, GATC sites with fewer than 60% of reads methylated on one or both strands were classified as hemimethylated or hypomethylated, respectively. Each caller identified different numbers of hypomethylated sites, and there is little overlap between the specific sites classified as hypomethylated by different callers (**Fig. 2B**). Across the 38,236 single-stranded GATC sites in the *E. coli* K-12 substr. MG1655 genome, Megalodon was the most precise, producing the fewest number of inconsistently classified GATC sites (162 sites, or 0.4%) outperforming Tombo, the second most consistent methylation caller with 294 inconsistent sites (**Fig. 2C; Supplementary Data**). DeepSignal and mCaller were the most inconsistent classifiers with 2340 and 3577 sites, respectively, inconsistently identified between sequencing replicates.

Given the high activity of Dam methyltransferase and the central role of GmATC in directing mismatch repair, GATC sites in *E. coli* are expected to be fully methylated except for a few key loci[10]. Thus, we compared how well each caller identified known undermethylated sites as a measure of caller accuracy. Previous estimates that measured 6mA by cutting with methylation-sensitive and -insensitive endonucleases have suggested that about 0.1% of *E. coli* GATC sites (or ~20 double-stranded GATC sites) are consistently found in a completely hypomethylated state *in vivo*[40]. In our analysis, Megalodon identified 22 consistently hypomethylated GATC sites across the three sequencing replicates. Tombo called 1 site as hypomethylated, while DeepSignal and mCaller identified 78 and 41 sites, respectively (**Table S1**). Based on these results, Megalodon is the most accurate as it identifies methylated sites at similar frequencies to those seen in previously published work[40].

As another measure of accuracy, we next examined how each caller identifies known undermethylated sites in the genome by focusing on the *srl* and *mtl* operons which are necessary for the import and metabolism of sorbitol and mannitol, respectively. Previous studies have established that in the absence of sorbitol and mannitol, transcriptional repressors bind to regions in the promoters that overlap with GATC motifs, preventing Dam methyltransferase binding and resulting in consistent undermethylation at these

GATC sites (**Table 1**)[37,40,50]. Of the four methylation callers, only mCaller and Megalodon correctly identified both sites as hypomethylated (**Fig. 2D**). Tombo identified the *srl* site as hypomethylated and the *mtl* site as hemimethylated, while DeepSignal called both sites as hemimethylated. Taken together, these data show that Megalodon is the most specific and accurate caller when classifying the methylation status of GATC sites using Nanopore data.

Methylation calling paired with CoMMA reveals variable 6mA methylation status between genomic features.

Once we determined that Megalodon was the most consistent and accurate methylation caller, we used the CoMMA "annotate" group of functions to characterize baseline genome-wide methylation trends for *E. coli* K-12 substr. MG1655 averaged for each site across the three sequencing replicates. CoMMA can use BED format files as input to annotate 6mA sites within various genomic regions. In this study, after the classification of 6mA methylation with Megalodon, all 6mA sites were annotated with data from RegulonDB, a curated database of *E. coli* K-12 transcriptional regulatory networks[51]. Methylation profiling revealed 177 double-stranded GATC sites are hypomethylated on at least one strand, of which 138 were hemimethylated and 39 were hypomethylated on both strands (**Table S2**). Intergenic 6mA sites were significantly less methylated compared to those within coding sequences (median percent methylated reads of 93% compared to 97%; Wilcoxon rank sum test, $P < 2.2\times10^{-16}$). This decrease at intergenic sites is likely due to the higher proportion of transcriptional regulatory regions and transcription factor binding sites in intergenic regions, where these sites may exclude Dam binding.

For a broader view of the factors that influence GATC methylation and to determine if hypomethylation was associated with GATC sites that may be shielded by protein binding, we looked at sites within intergenic regions where DNA-protein interactions likely occur, such as sigma factor binding sites and binding sites of other transcriptional regulatory proteins[10,36,37,52,53]. Overall, within these intergenic regions, there is a general decrease in methylation around promoters. GATC sites within -10 and -35 promoter elements were significantly less methylated (95.2% for -10 elements, $P_{-10} = 6.7\times10^{-8}$; 90.4% for -35 elements, $P_{-35} =$

6.7×10^{-10}, Wilcoxon signed rank test with Benjamini-Hochberg correction) than the genome-wide median percent methylation (96.6%). This is best observed when methylation is averaged across all promoter regions in the genome; there is a lower frequency of methylated reads around 100 bp before and after the transcription start site, with a maximum decrease in methylation of ~5% centered on the -35 element (**Fig 3A**). Moreover, -35 promoter elements known to be bound by σ^{70}, σ^{38}, and σ^{24} had significantly lower percent methylation than the genome-wide median (Wilcoxon signed rank test with BH correction: $P_{sigma70} = 2.0\times10^{-10}$; $P_{sigma38} = 7.0\times10^{-4}$; $P_{sigma24} = 1.6\times10^{-5}$) (**Fig 3B**).

In addition to core promoter elements, binding sites of other transcriptional regulatory proteins within intergenic regions also exhibited differences in their degree of hypomethylation (**Fig 3C**). Specifically, the global regulators CRP, Fis, Fnr, and Nac were significantly less methylated than the genome-wide median (Wilcoxon signed rank test with BH correction: $P_{CRP} = 2.0\times10^{-18}$; $P_{Fis} = 4.2\times10^{-3}$; $P_{Fnr} = 9.1\times10^{-6}$; $P_{Nac} = 2.3\times$). Differences in methylation were also observed within coding regions for a few genomic features (**Fig 3C**). Across the 12 cryptic prophages in the *E. coli* K-12 genome, significantly decreased methylation is observed in 9 of them: CP4-44, CP4-57, CP4-6, CPS-53, DLP12, e14, Qin, Rac, and PR-Y (CP4-44: 94.0% methylation, $P = 6.2\times10^{-4}$; CP4-57: 92.3%, $P = 9.4\times10^{-16}$; CP4-6: 95.3%, $P = 2.5\times10^{-13}$; CPS-53: 92.4%, $P = 4.7\times10^{-6}$; DLP12: 88.5%, $P = 3.7\times10^{-13}$; e14: 92.1%, $P = 1.1\times10^{-8}$; Qin: 95.1%, $P = 2.9\times10^{-6}$; Rac: 94.9%, $P = 4.3\times10^{-8}$; PR-Y: 93.8%, $P = 2.6\times10^{-2}$; Wilcoxon signed rank test with BH correction)[54,55]. Of the various IS elements, *IS*150, *IS*4, *IS*30A, *IS*30C, and *IS*30D had significantly decreased methylation (*IS*150: 89.4% methylation, $P = 2.6\times10^{-3}$; *IS*4: 94.1%, $P = 1.3\times10^{-2}$; *IS*30A: 89.7%, $P = 4.3\times10^{-2}$; *IS*30C: 85.8%, $P = 1.5\times10^{-2}$; *IS*30D: 86.7%, $P = 2.2\times10^{-2}$)[56–58]. Lastly, some genomic features, such as REP elements and terminator sequences, do not show measurable differences in 6mA methylation compared to the genome-wide average. However, these regions either do not contain many 6mA sites or are poorly annotated. More sequencing replicates and better genomic annotations could allow for the detection of higher-resolution methylation differences and the characterization of methylation on abundant but unannotated sites.

Differential methylation with CoMMA shows MMR- clones evolve decreased GATC methylation.

After characterizing a baseline *E. coli* K-12 substr. MG1655 6mA methylome, we used the CoMMA "compare" functions to compare GATC methylation across the experimental ancestor PMF2 strain and eight isolated clones from populations that were experimentally evolved for 2400 generations with either intact (WT) or disrupted mismatch repair (MMR-)[59]. Briefly, CoMMA adapts differential methylation package methylKit to identify 6mA sites with significant differences in methylation status between bacterial samples[60]. For a broad view of how methylation evolves, we focused on two populations from each of the WT or MMR- genetic backgrounds; from each of these populations, we selected two clones and ensured each clone belonged to a different evolved subpopulation. This sampling scheme would allow us to evaluate how methylation evolves within and among populations depending on their ability to conduct methyl-directed mismatch repair.

As with the ancestor, experimentally evolved clones were Nanopore sequenced, and methylated sites were called using Megalodon. The median genome-wide methylation percentage of both genetic backgrounds decreased during experimental evolution with MMR- lines exhibiting a significantly larger median decrease in methylation (-0.15%) compared to evolved WT lines (-0.03%; Wilcoxon rank sum test, $P < 2.2\times10^{-16}$, **Fig. S4**). Next, we used CoMMA to determine which specific GATC sites were differentially methylated between the evolved lines and the ancestor (**Fig 4A-B**). Using cutoffs $P < 0.05$ and percent methylation change >10%, there were significantly more differentially methylated sites in the MMR- lines (28 sites with increased methylation, 141 sites with decreased methylation) compared to the evolved WT lines (22 increased, 16 decreased; Fisher's exact test, $P < 2.2\times10^{-16}$).

Both evolved backgrounds also had a significant overrepresentation of differentially methylated sites within certain genomic regions, especially in sites where DNA-protein interactions occur. Differentially methylated sites in both evolved backgrounds were significantly over-enriched in transcription factor binding motifs and IS elements (Fisher's exact test with FDR correction, $P < 0.05$, **Table 2**). Sites with increased methylation in both evolved backgrounds were also over-enriched within cryptic prophages, and sites with decreased methylation were over-enriched in Rho-independent terminators. The MMR- evolved lines also showed enrichment for under-methylated sites within cryptic prophages ($P < 0.0001$) and within promoters, specifically σ^{70}-binding sites in the -35 region ($P < 0.0001$).

Comparison of the full 6mA methylomes between evolved clones and their ancestor revealed a non-deterministic change in 6mA in the absence of MMR consistent with random genetic drift. The percent of methylated reads at all GATC sites was compared between samples by clustering using non-metric multidimensional scaling (NMDS; **Fig. 4C**). The evolved WT clones and their ancestor clustered together, showing that their 6mA methylomes remain similar after extended evolution (Pairwise PERMANOVA Ancestor vs. WT, P = 0.168; Analysis of multivariate homogeneity of group dispersions, ANOVA with Tukey's HSD: $P_{Ancestor\text{-}WT} = 0.99$). In contrast, MMR- evolved clones diverge from the ancestor in random directions, and the MMR- clones do not cluster together. The centroid of the MMR- clones was near that of the ancestor and WT clones but significantly different than the WT clones, and MMR- clone dispersions were significantly greater than both the ancestor and WT groups (Pairwise PERMANOVA WT vs. MMR-, P = 0.026; Analysis of multivariate homogeneity of group dispersions, ANOVA with Tukey's HSD: $P_{Ancestor\text{-}MMR\text{-}} = 0.018$, $P_{WT\text{-}MMR\text{-}} = 0.010$). This suggests that 6mA differences that occur during evolution in MMR- lines are non-deterministic, resulting in significant changes to methylation but not strongly in any one direction. MMR- clones from the same population did not cluster together, indicating that shared evolutionary history does not have a large effect on broad 6mA differences over this time scale. Together, the significant number of under-methylated sites and the random divergence of differentially methylated sites in the MMR- clones compared to the ancestor suggest that 6mA is shaped by genetic drift in the absence of MMR.

MMR- evolved lines accumulate GATC mutations at a rate that indicates relaxed selection.

After identifying decreased maintenance of 6mA in the absence of MMR with CoMMA, we sought to genomically identify whether GATC sites in MMR- clones were also under decreased evolutionary constraint. Here, we characterized the fraction of mutations that altered GATC sites across the experimentally evolved clones from both evolved genetic backgrounds. After ~2400 generations, MMR- clones had acquired a total of 959 mutations with 34 mutations (3.5% of mutations) causing a gain or loss of a GATC site (**Table 3, Table S3**). In contrast, the WT clones accumulated a total of 33 mutations, none of which had affected GATC sites. Since the relative lack of mutations in WT clones limited our statistical power, we compared the mutation

rates of the experimentally evolved clones to mutation rates calculated in a minimally-selective mutation accumulation (MA) study by Lee *et al*.[43]. As this study directly measured the mutation rates of the same WT and MMR- PFM2 strain, we are able to accurately determine the expected rate of GATC mutation in the absence of selection[43].

Reanalysis of the MA data to calculate mutation rates in tetranucleotide contexts determined that 3.5% of mutations that arose in MMR- lines during MA were in GATC sites, and is equal to the fraction of mutations affecting GATC sites in the MMR- lines during experimental evolution[43]. The similarity of these GATC mutation rates suggests that experimentally evolved MMR- clones acquire mutations in GATC sites at the neutral rate and may be evolving under relaxed selection. Additional evidence that GATC sites and associated adenine methylation may be evolving under relaxed selection in experimentally evolved MMR- clones include that the clones isolated from the MMR- Population 113 both acquired mutations resulting in a nonsynonymous D20A substitution proximal to the GATC recognition domain of *dam* methyltransferase[61]. While the exact consequences of this mutation have not been examined, this substitution could affect DNA binding and GATC recognition, indicating that perfect Dam function may not be necessary in the absence of MMR. Together, these data suggest that MMR is a major factor in maintaining GATC sites and *dam* methyltransferase, and in the absence of MMR these loci are vulnerable to mutation.

DISCUSSION

IIn this study, we present CoMMA, a pipeline for annotating and identifying differentially methylated 6mA sites in bacterial samples using Nanopore sequencing data. We evaluated the performance of four 6mA methylation callers and found that Megalodon was the most reliable for identifying methylated adenines at GATC sites. This is consistent with recent studies that have shown Megalodon to be one of the most accurate programs for identifying 5mC modifications from Nanopore data[5]. We then used Megalodon and CoMMA to produce a comprehensive genome-wide analysis of 6mA methylation in *E. coli* K-12 substr. MG1655 using long-read sequencing data. Our analysis shows that Nanopore sequencing and pipelines such as CoMMA now provide reliable and cost-effective methods for characterizing methylation across different motifs. To our

knowledge, this is the first study of its kind to produce a high-resolution characterization of the *E. coli* K-12 MG1655 6mA methylome that includes a functional and genic classification of percent 6mA methylation. Patterns of reduced methylation at intergenic regions, promoter elements, DNA binding sites, and cryptic prophages could be the result of the physical exclusion of Dam methyltransferase at these sites and could have downstream effects on gene regulation and stress response. The specific effects of these methylation differences should be investigated in the future with transcriptomic and biochemical experiments.

We also applied the CoMMA pipeline to investigate the effect of ~2400 generations of evolution on WT and MMR- *E. coli* MG1655 clones. We found that the MMR- clones showed a significant decrease in methylation, indicating that 6mA is to some degree dispensable in the absence of MMR. While methylation was largely stable at the majority of GATC sites, MMR- clones showed significantly more differentially methylated sites compared to the WT. Interestingly, these differentially methylated sites are more frequently found at DNA binding sites. Adenine modification can affect specific binding affinity where 6mA coincides with a DNA binding site[62,63]. Thus, without the genome-wide pressure to maintain 6mA to minimize mutation rates, MMR- clones are free to alter 6mA at these sites. Most of these sites likely display decreased methylation due to binding competition between transcription factors and the methyltransferase. Altered 6mA levels could then affect transcription factor binding and further alter cellular phenotypes. Our study demonstrates that differential methylomics is a powerful approach that can directly resolve differences between clones that may otherwise not be inferred by genetic differences in an experiment of this time scale. The CoMMA pipeline is thus a useful tool that is complementary to genetic approaches.

One interesting aspect of our study is the observation that one of the MMR- evolved populations (P113) acquired a mutation in the *dam* adenine methyltransferase resulting in a single amino acid substitution (DamD20A). The consequence of this mutation is unclear, although this residue is proximal to three S-adenosylmethionine-binding sites at K10, K14 and D54, and could potentially affect protein folding[61]. While there are no broad methylation differences within these clones compared to others of the same background, it is possible that it may have contributed to the observed changes in methylation in these clones. Further

studies are needed to investigate if this mutation altered Dam activity and how this affects genome-wide methylation patterns.

One potential impact of our findings is on the mutational spectrum and adaptive landscape of bacteria. DNA base modifications can affect the rates of spontaneous deamination, such as 5mC which is frequently deaminated to thymine[64–66]. Similarly, adenine and 6mA may be deaminated to hypoxanthine, which pairs to cytosine during replication[67,68]. While the rates of spontaneous deamination in adenine versus 6mA have not been studied, a recent study found a correlation between 6mA and A→G transitions in *Neisseria meningitidis*[69]. Thus, decreases in 6mA after MMR loss may result in an increased A→G transition rate. This change to the mutational spectrum could potentially affect how MMR- clones to adapt to changing environments and may influence the trajectory of their evolutionary paths.

This study is one of the first-ever analyses of differential methylation between evolved bacterial lineages at this time scale. The CoMMA workflow could be used to identify changes to 6mA at GATC sites in any bacterial population, and it could be modified for use in other methylation motifs. Third-generation sequencing and pipelines such as CoMMA create opportunities for the seamless integration of genetic and epigenetic characterization of an organism, especially in the context of 6mA methylation, which was previously difficult to characterize[36,53,70,71]. These tools are now sufficiently reliable and high-throughput to powerfully analyze genetic and epigenetic evolution in bacteria. The evolutionary consequences of bacterial epigenetics are largely unstudied. Recent studies have shown the significant effects that bacterial DNA methylation can have on gene expression, and with our method, such analyses can be extended through time or between bacterial isolates[45,46,72,73]. These tools can be used to elucidate the relationship between mutation, epimutation, and gene expression as it relates to adaptation to challenging environments.

METHODS

Bacterial strains and culture conditions

All bacterial strains were cultured and sequenced to identify mutations as part of a long-term experimental evolution study that has been described previously[59]. Briefly, the WT ancestor strain PFM2 is a

prototrophic derivative of *E. coli* K-12 MG1655, and the MMR-deficient strain PFM5 is derived from PFM2 with an in-frame deletion of the *mutL* gene[59]. Long-term cultures were started by inoculating 10 mL of LB-Miller broth (BD Difco) in 16- x 100-mm glass culture tubes with a single colony of the progenitor strain[59]. Cultures were maintained shaking at 175 rpm, alternating transfers after 24 h of growth at 37 °C and 48 h of growth at 25 °C. Cultures of interest in this study were transferred by thoroughly vortexing the culture tube and inoculating 9 mL of LB-Miller broth with 1 mL of culture, resulting in a transfer of about 10^9 cells[59]. Populations were grown and transferred for 2000 total generations of growth. After 2000 generations, samples from the evolving populations were streaked for isolation and eight random clones were saved in 40% glycerol at -80 °C for further and future experimentation.

Nucleic acid isolation and Nanopore sequencing

Evolved and WT ancestor clones were revived from frozen storage on LB agar by incubating overnight at 37 °C. For each clone, a single isolated colony was randomly selected and restreaked on to fresh LB agar and again incubated overnight at 37 °C for DNA extraction. Following overnight incubation, colonies were collected from the agar using a 10mL loop and inoculated directly into DNA extraction buffer and DNA was further isolated with the UltraClean Microbial DNA Kit (Qiagen). Following recommendations from ONT, Extracted DNA was assessed for quantity using the Qubit High Sensitivity dsDNA kit and quality using the Genomic DNA ScreenTape on a TapeStation 4200 (Agilent). Libraries were then prepared for Nanopore sequencing using the PCR-free Rapid DNA sequencing Kit (SQK-RAD004, Nanopore) paired with the Rapid Barcoding Kit (SQK-RBK004). Prepared libraries were pooled and loaded on to a R.9.4.1 flow cell and sequenced on a MinION for 48h with live basecalling in MinKNOW deactivated.

Bioinformatics and sequencing analysis

To determine precision of methylation calling, we compared results for each of four methylation calling pipelines across three Nanopore sequencing replicates of the WT ancestor PFM2 and determined the number of inconsistently called sites. Initial Nanopore basecalling by converting fast5 files into fastq was conducted

using GuppyGPU (v. 3.1.5+781ed57). Following basecalling, samples were demultiplexed with Guppy GPU (Supplementary Table XX). Average read length across all samples is 5583 +/- 275 nucleotides (+/- s.e.m). Following basecalling and demultiplexing, we evaluated four different pipelines (Tombo (v.1.5), mCaller (v.0.3), Deepsignal (v.0.1.6), and Megalodon (v.0.1.0)) for their precision and accuracy to identify base modifications and determine methylated adenines within GATC sites[4,6,47,49]. Both Tombo and Megalodon were created by ONT for methylation calling, with Tombo being an older release that relies on statistical methods to annotate non-canonical bases and Megalodon being a newer release based on machine learning with deep neural networks[47,49]. To increase accuracy, Megalodon can be run with a VCF file containing the position and identity of known variants, for which we used Illumina sequencing data for evolved clones that was previously published[55]. For pipelines not developed by ONT: DeepSignal applies a deep neural network to data that has been preprocessed by Tombo; and mCaller applies a deep neural network to data that has been preprocessed by Nanopolish. When running the mCaller pipeline we used Nanopolish v. 0.11.1). Across all samples, the mean read coverage across GATC sites was 56 reads per site (**Fig. S2**).

Identification of mutations affecting GATC sites in evolved clones was conducted using previously published WGS Illumina sequencing data. To identify the rate of mutations affecting GATC sites under minimal selection, we utilized previously published mutation accumulation data for the *E. coli* strains that served as the experimental ancestor for the WT (PFM2) and MMR- (PMF5, PMF2:Δ*mutL*)[43]. For both the experimentally evolved clones and the MA lines, mutations were called using GATK best practices and the genic location of mutations was annotated using SnpEff v. 4.3 by comparing VCF files of identified mutations for each clone to the *Escherichia coli* K-12 str. MG1655 reference genome (INSDC accession: U00096.3)[74]. To confirm a mutation occurred in a GATC site, the tetranucleotide context of each mutation was identified with the fill-fs -l 5 function from vcftools (v. 0.1.12) to annotate each variant with the flanking five upstream and downstream nucleotides[75].

GATC sites were annotated as within different genomic features using all combined databases in RegulonDB v. 10.9[76]. These annotations were concatenated into a BED file, which was used to annotate GATC sites with the CoMMA function annotateMethylSites.

Differential methylation calling

Differential methylation analysis was performed in R using CoMMA, a package that serves as a wrapper for methylKit v. 3.13[60]. GATC sites that had acquired mutations in any clone were removed from differential methylation analysis so that only GATC sites shared between samples were considered. To ensure the reliability of differential methylation analysis, CoMMA includes the following modifications to the standard methylKit workflow: 1) any sites covered by fewer than 10 reads were excluded from analysis; 2) methylation coverage at each site was normalized by median coverage; and 3) sites were ultimately considered to be differentially methylated based on a q-value cutoff of < 0.1 and percent methylation difference > 10%. Differential methylation was called separately between the ancestor and either the evolved MMR- or WT clones. Differential methylation was calculated by logistic regression with a sliding linear model (SLIM) to correct for multiple hypothesis testing[60,77].

Statistical analysis and ordination

Statistical analysis comparing percent methylation (or -value) was performed on M-values to remove the heteroscedasticity observed in -value distributions[78]. M-values were calculated as shown in Equation 1:

$$M = log2\left(\frac{\beta}{1-\beta+0.001}\right) \quad (1)$$

All statistical analysis was done in R. M-values of GATC sites within different genomic features were compared using the Kruskal-Wallis rank sum test, and each feature was compared to the genome-wide median M-value using a one-sample Wilcoxon signed rank test with Benjamini-Hochberg correction for multiple comparisons[79].

Ordination of samples was performed after normalizing -values. Since -values are dependent on coverage, samples with low coverage stand out when visualized with non-metric ordination methods. Therefore, before ordination, we normalized coverage and methylated read counts between samples to reduce the effects of coverage on ordination distance. First, a linear scaling equation for each sample was found by normalizing coverage distributions of all samples using quantile normalization[80]. Each sample's

scaling equation was calculated as the simple linear regression line between sample coverage and normalized coverage. This scaling equation was then applied to each sample's coverage and methylated read count, both of which were used to calculate normalized -values. Ordination of samples was performed using the R package vegan v. 2.6-4[81]. Non-metric multidimensional scaling (NMDS) was used to visualize sample clustering based on a Euclidean distance matrix calculated from normalized -values.

Data and code availability

Raw sequencing data are available for download at XXXXXX. Code and data are available at https://github.com/carl-stone/CoMMA.

REFERENCES

1. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
2. Davis, B. M., Chao, M. C. & Waldor, M. K. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Curr. Opin. Microbiol.* **16**, 192–198 (2013).
3. Korlach, J. & Turner, S. W. Going beyond five bases in DNA sequencing. *Curr. Opin. Struct. Biol.* **22**, 251–261 (2012).
4. McIntyre, A. B. R. *et al.* Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. *Nat. Commun.* **10**, 579 (2019).
5. Yuen, Z. W.-S. *et al.* Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nat. Commun.* **12**, 3438 (2021).
6. Ni, P. *et al.* DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics* **35**, 4586–4595 (2019).
7. Liu, Q. *et al.* Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nat. Commun.* **10**, 2449 (2019).
8. Jones, P. A. & Baylin, S. B. The Epigenomics of Cancer. *Cell* **128**, 683–692 (2007).
9. Sánchez-Romero, M. A. & Casadesús, J. The bacterial epigenome. *Nat. Rev. Microbiol.* **18**, 7–20 (2020).
10. Wion, D. & Casadesús, J. N6-methyl-adenine: an epigenetic signal for DNA-protein interactions. *Nat. Rev. Microbiol.* **4**, 183–192 (2006).
11. Seib, K. L., Srikhanta, Y. N., Atack, J. M. & Jennings, M. P. Epigenetic Regulation of Virulence and Immunoevasion by Phase-Variable Restriction-Modification Systems in Bacterial Pathogens. *Annu. Rev. Microbiol.* **74**, 655–671 (2020).
12. Adhikari, S. & Curtis, P. D. DNA methyltransferases and epigenetic regulation in bacteria. *FEMS Microbiol. Rev.* **40**, 575–591 (2016).

13. Ardissone, S. Cell Cycle Constraints and Environmental Control of Local DNA Hypomethylation in α-Proteobacteria. *PLOS Genet* (2016) doi:10.1371/journal.pgen.1006499.

14. Reisenauer, A., Kahng, L. S., McCollum, S. & Shapiro, L. Bacterial DNA Methylation: a Cell Cycle Regulator? *J. Bacteriol.* **181**, 5135–5139 (1999).

15. Collier. A DNA methylation ratchet governs progression through a bacterial cell cycle. *Proc Natl Acad Sci USA* (2007) doi:10.1073/pnas.0708112104.

16. Sánchez-Romero, M. A., Cota, I. & Casadesús, J. DNA methylation in bacteria: from the methyl group to the methylome. *Curr. Opin. Microbiol.* **25**, 9–16 (2015).

17. De Ste Croix, M. *et al.* Phase-variable methylation and epigenetic regulation by type I restriction-modification systems. *FEMS Microbiol. Rev.* **41**, S3–S15 (2017).

18. Nye, T. M. *et al.* DNA methylation from a Type I restriction modification system influences gene expression and virulence in Streptococcus pyogenes. *PLOS Pathog.* **15**, e1007841 (2019).

19. Oliveira, P. H. & Fang, G. Conserved DNA Methyltransferases: A Window into Fundamental Mechanisms of Epigenetic Regulation in Bacteria. *Trends Microbiol.* **29**, 28–40 (2021).

20. Geier, G. E. & Modrich, P. Recognition sequence of the dam methylase of Escherichia coli K12 and mode of cleavage of Dpn I endonuclease. *J. Biol. Chem.* **254**, 1408–1413 (1979).

21. Herman, G. E. & Modrich, P. Escherichia coli dam methylase. Physical and catalytic properties of the homogeneous enzyme. *J. Biol. Chem.* **257**, 2605–2612 (1982).

22. Marinus, M. G. & Morris, N. R. Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12. *J. Bacteriol.* **114**, 1143–1150 (1973).

23. Blattner, F. R. *et al.* The Complete Genome Sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).

24. Boye, E. & Løbner-Olesen, A. The role of dam methyltransferase in the control of DNA replication in E. coli. *Cell* **62**, 981–989 (1990).

25. Pukkila, P. J., Peterson, J., Herman, G., Modrich, P. & Meselson, M. Effects of High Levels of DNA Adenine Methylation on Methyl-Directed Mismatch Repair in *Escherichia coli*. *Genetics* **104**, 571–582 (1983).

26. Schlagman, S. L., Hattman, S. & Marinus, M. G. Direct role of the Escherichia coli Dam DNA methyltransferase in methylation-directed mismatch repair. *J. Bacteriol.* **165**, 896–900 (1986).

27. Messer, W. & Noyer-Weidner, M. Timing and targeting: The biological functions of Dam methylation in E. coli. *Cell* **54**, 735–737 (1988).

28. Messer, W., Bellekes, U. & Lother, H. Effect of dam methylation on the activity of the E. coli replication origin, oriC. *EMBO J.* **4**, 1327–1332 (1985).

29. Herman, G. E. Escherichia coli K-12 clones that overproduce dam methylase are hypermutable. *J Bacteriol* (1981) doi:10.1128/jb.145.1.644-646.1981.

30. Lahue, R. S. & Modrich, P. Methyl-directed DNA mismatch repair in Escherichia coli. *Mutat. Res. Mol. Mech. Mutagen.* **198**, 37–43 (1988).

31. Roberts, D., Hoopes, B. C., McClure, W. R. & Kleckner, N. IS10 transposition is regulated by DNA adenine methylation. *Cell* **43**, 117–130 (1985).

32. Tomcsanyi, T. & Berg, D. E. Transposition effect of adenine (Dam) methylation on activity of O end mutants of IS50. *J. Mol. Biol.* **209**, 191–193 (1989).

33. Torreblanca, J. & Casadesús, J. DNA adenine methylase mutants of Salmonella typhimurium and a novel dam-regulated locus. *Genetics* **144**, 15–26 (1996).

34. Heithoff, D. M., Sinsheimer, R. L., Low, D. A. & Mahan, M. J. An essential role for DNA adenine methylation in bacterial virulence. *Science* **284**, 967–970 (1999).

35. Lu. SeqA: A negative modulator of replication initiation in E. coli. *Cell* (1994) doi:10.1016/0092-8674(94)90156-2.

36. Sánchez-Romero, M. A. & Casadesús, J. Waddington's Landscapes in the Bacterial World. *Front. Microbiol.* **12**, 685080 (2021).

37. van der Woude, M., Hale, W. B. & Low, D. A. Formation of DNA Methylation Patterns: Nonmethylated GATC Sequences in gut and papOperons. *J. Bacteriol.* **180**, 5913–5920 (1998).

38. Hernday, A. D., Braaten, B. A. & Low, D. A. The Mechanism by which DNA Adenine Methylase and PapI Activate the Pap Epigenetic Switch. *Mol. Cell* **12**, 947–957 (2003).

39. Hernday, A. D., Braaten, B. A., Broitman-Maduro, G., Engelberts, P. & Low, D. A. Regulation of the Pap Epigenetic Switch by CpxAR: Phosphorylated CpxR Inhibits Transition to the Phase ON State by Competition with Lrp. *Mol. Cell* **16**, 537–547 (2004).

40. Tavazoie, S. & Church, G. M. Quantitative whole-genome analysis of DNA-protein interactions by in vivo methylase protection in E. coli. *Nat. Biotechnol.* **16**, 566–571 (1998).

41. Li, G.-M. Mechanisms and functions of DNA mismatch repair. *Cell Res.* **18**, 85–98 (2008).

42. Lu, A. L., Clark, S. & Modrich, P. Methyl-directed repair of DNA base-pair mismatches in vitro. *Proc. Natl. Acad. Sci. U. S. A.* **80**, 4639–4643 (1983).

43. Lee, H., Popodi, E., Tang, H. & Foster, P. L. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci.* **109**, E2774–E2783 (2012).

44. Posnick, L. M. & Samson, L. D. Influence of S-Adenosylmethionine Pool Size on Spontaneous Mutation, Dam Methylation, and Cell Growth of Escherichia coli. *J. Bacteriol.* **181**, 6756–6762 (1999).

45. Westphal, L. L., Sauvey, P., Champion, M. M., Ehrenreich, I. M. & Finkel, S. E. Genomewide Dam Methylation in Escherichia coli during Long-Term Stationary Phase. *mSystems* **1**, e00130-16 (2016).

46. Walworth, N. G. *et al.* Long-Term m5C Methylome Dynamics Parallel Phenotypic Adaptation in the Cyanobacterium Trichodesmium. *Mol. Biol. Evol.* **38**, 927–939 (2021).

47. Oxford Nanopore Technologies. Megalodon. (2022).

48. Oxford Nanopore Technologies. Tombo.

49. Oxford Nanopore Technologies. Tombo Documentation Modified Base Detection.

https://nanoporetech.github.io/tombo/modified_base_detection.html.

50. Casadesús, J. & Low, D. Epigenetic Gene Regulation in the Bacterial World. *Microbiol. Mol. Biol. Rev.* **70**, 830–856 (2006).

51. Tierrafría, V. H. *et al.* RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in Escherichia coli K-12. *Microb. Genomics* **8**, (2022).

52. Wang, M. X. & Church, G. M. A whole genome approach to in vivo DNA-protein interactions in E. coli. *Nature* **360**, 606–610 (1992).

53. Sanchez-Romero. Contribution of DNA adenine methylation to gene expression heterogeneity in Salmonella enterica. *Nucleic Acids Res* (2020) doi:10.1093/nar/gkaa730.

54. Wang, X. *et al.* Cryptic prophages help bacteria cope with adverse environments. *Nat. Commun.* **1**, 147 (2010).

55. Ramisetty, B. C. M. & Sudhakari, P. A. Bacterial 'Grounded' Prophages: Hotspots for Genetic Renovation and Innovation. *Front. Genet.* **10**, (2019).

56. Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.* **38**, 865–891 (2014).

57. Lee, H., Doak, T. G., Popodi, E., Foster, P. L. & Tang, H. Insertion sequence-caused large-scale rearrangements in the genome of *Escherichia coli*. *Nucleic Acids Res.* gkw647 (2016) doi:10.1093/nar/gkw647.

58. Naas, T. Insertion sequence-related genetic variation in resting Escherichia coli K-12. *Genetics* (1994) doi:10.1093/genetics/136.3.721.

59. Behringer, M. G. *et al.* Escherichia coli cultures maintain stable subpopulation structure during long-term evolution. *Proc. Natl. Acad. Sci.* **115**, E4642–E4650 (2018).

60. Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87 (2012).

61. Horton, J. R., Liebert, K., Bekes, M., Jeltsch, A. & Cheng, X. Structure and Substrate Recognition of the Escherichia coli DNA Adenine Methyltransferase. *J. Mol. Biol.* **358**, 559–570 (2006).

62. Maier, J. A. H. & Jeltsch, A. Design and Application of 6mA-Specific Zinc-Finger Proteins for the Readout of DNA Methylation. in *Zinc Finger Proteins: Methods and Protocols* (ed. Liu, J.) 29–41 (Springer, 2018). doi:10.1007/978-1-4939-8799-3_3.

63. Sera, T. & Uranga, C. Rational design of artificial zinc-finger proteins using a nondegenerate recognition code table. *Biochemistry* **41**, 7074–7081 (2002).

64. Nabel, C. S., Manning, S. A. & Kohli, R. M. The Curious Chemical Biology of Cytosine: Deamination, Methylation,and Oxidation as Modulators of Genomic Potential. *ACS Chem. Biol.* **7**, 20–30 (2012).

65. Shapiro, R. & Klein, R. S. The Deamination of Cytidine and Cytosine by Acidic Buffer Solutions. Mutagenic Implications*. *Biochemistry* **5**, 2358–2362 (1966).

66. Lindahl, T. & Nyberg, B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13**, 3405–3410 (1974).

67. O'Brown, Z. K. & Greer, E. L. N6-Methyladenine: A Conserved and Dynamic DNA Mark. in *DNA Methyltransferases - Role and Function* (eds. Jeltsch, A. & Jurkowska, R. Z.) 213–246 (Springer International Publishing, 2016). doi:10.1007/978-3-319-43624-1_10.

68. Basilio, C., Wahba, A. J., Lengyel, P., Speyer, J. F. & Ochoa, S. Synthetic polynucleotides and the amino acid code, V*. *Proc. Natl. Acad. Sci.* **48**, 613–616 (1962).

69. Sater, M. R. A. *et al.* DNA Methylation Assessed by SMRT Sequencing Is Linked to Mutations in Neisseria meningitidis Isolates. *PLOS ONE* **10**, e0144612 (2015).

70. Payelleville, A. & Brillard, J. Novel Identification of Bacterial Epigenetic Regulations Would Benefit From a Better Exploitation of Methylomic Data. *Front. Microbiol.* **12**, (2021).

71. Marinus, M. G. & Casadesus, J. Roles of DNA adenine methylation in host–pathogen interactions: mismatch repair, transcriptional regulation, and more. *FEMS Microbiol. Rev.* **33**, 488–503 (2009).

72. Brunet. Fur-dam regulatory interplay at an internal promoter of the enteroaggregative Escherichia coli Type VI secretion sci1 gene cluster. *J Bacteriol* (2020) doi:10.1128/JB.00075-20.

73. Mehling. A Dam methylation mutant of Klebsiella pneumoniae is partially attenuated. *FEMS Microbiol Lett.* (2007) doi:10.1111/j.1574-6968.2006.00581.x.

74. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, 80–92 (2012).

75. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

76. Santos-Zavaleta, A. *et al.* RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Res.* **47**, D212–D220 (2019).

77. Wang, H.-Q., Tuominen, L. K. & Tsai, C.-J. SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinforma. Oxf. Engl.* **27**, 225–231 (2011).

78. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).

79. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

80. Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).

81. Oksanen, J. *et al. vegan: Community Ecology Package*. (2022).

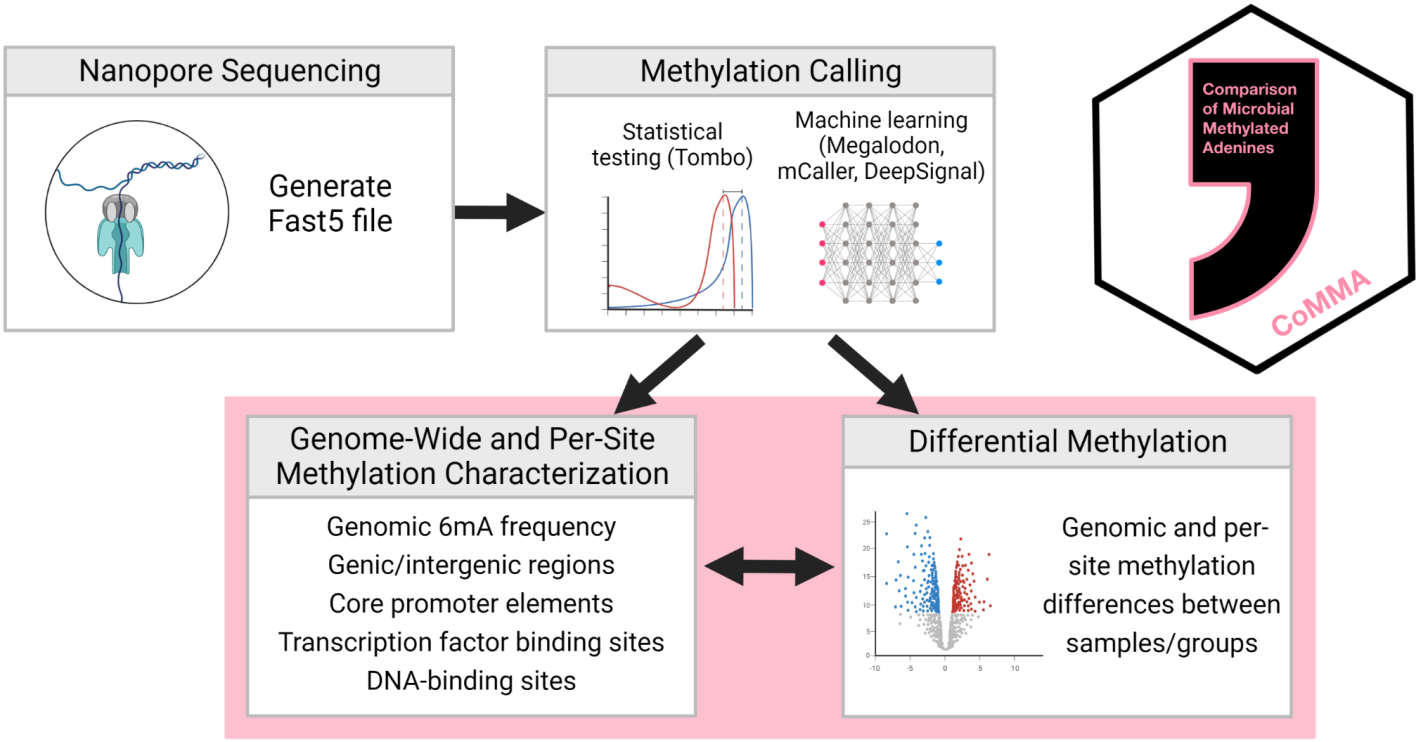

Figure 1. Summary of the CoMMA pipeline.

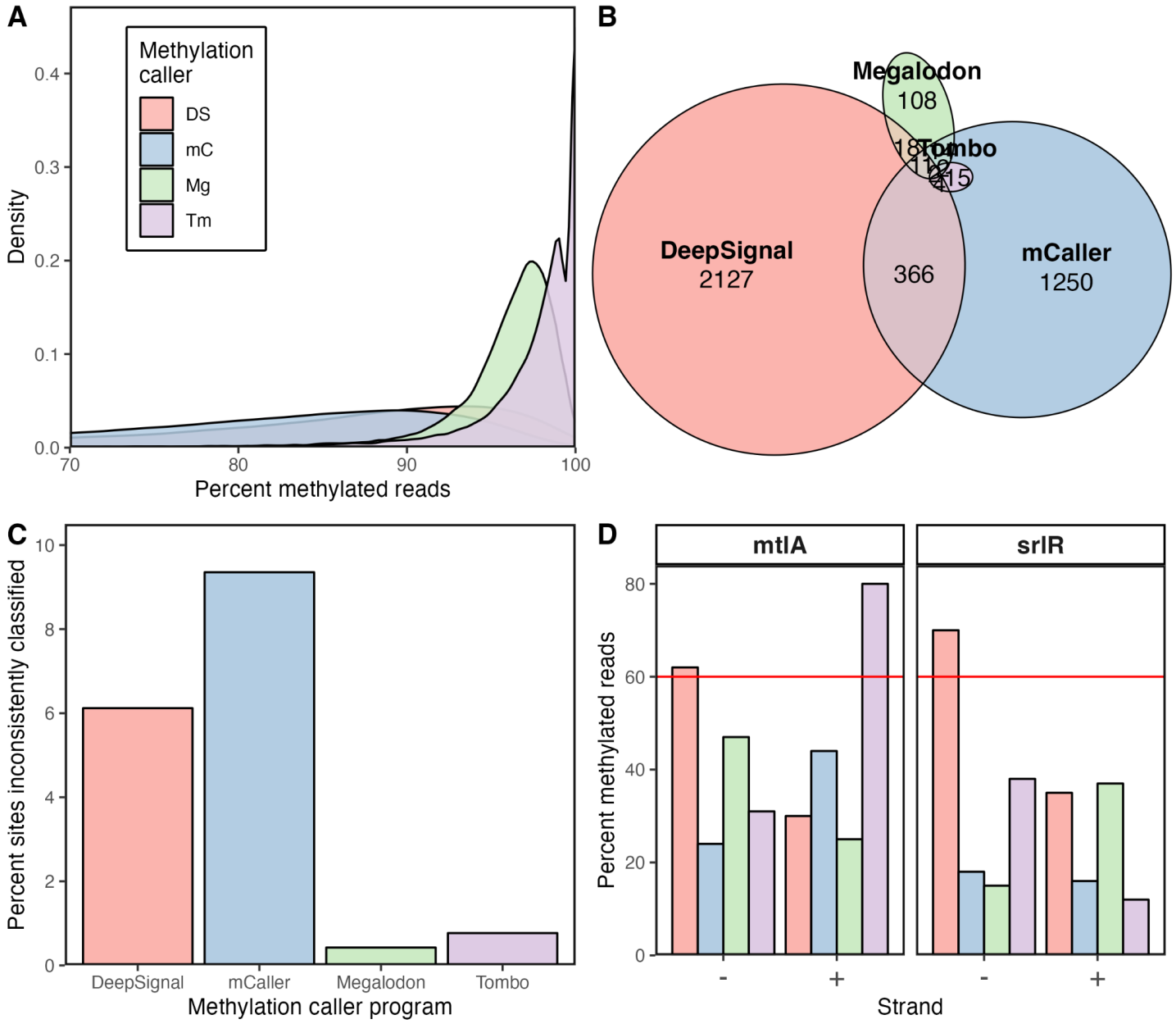

Figure 2. Comparison of bioinformatics programs on identifying methylated reads at GATC sites. (A) Distribution of percent of reads classified as methylated at all GATC sites in the genome for each caller. Dashed lines represent median percent methylation across all GATC sites for, from left to right, mCaller (83%), DeepSignal (87%), Megalodon (97%), and Tombo (99%). (B) Overlap between GATC sites classified as undermethylated by each program. Each area is labeled with the number of GATC sites. (C) Percentage of methylated GATC sites inconsistently classified as methylated across WT replicates for each program. (D) Percentage of reads classified as methylated by each program at GATC sites previously reported as <60% methylated.

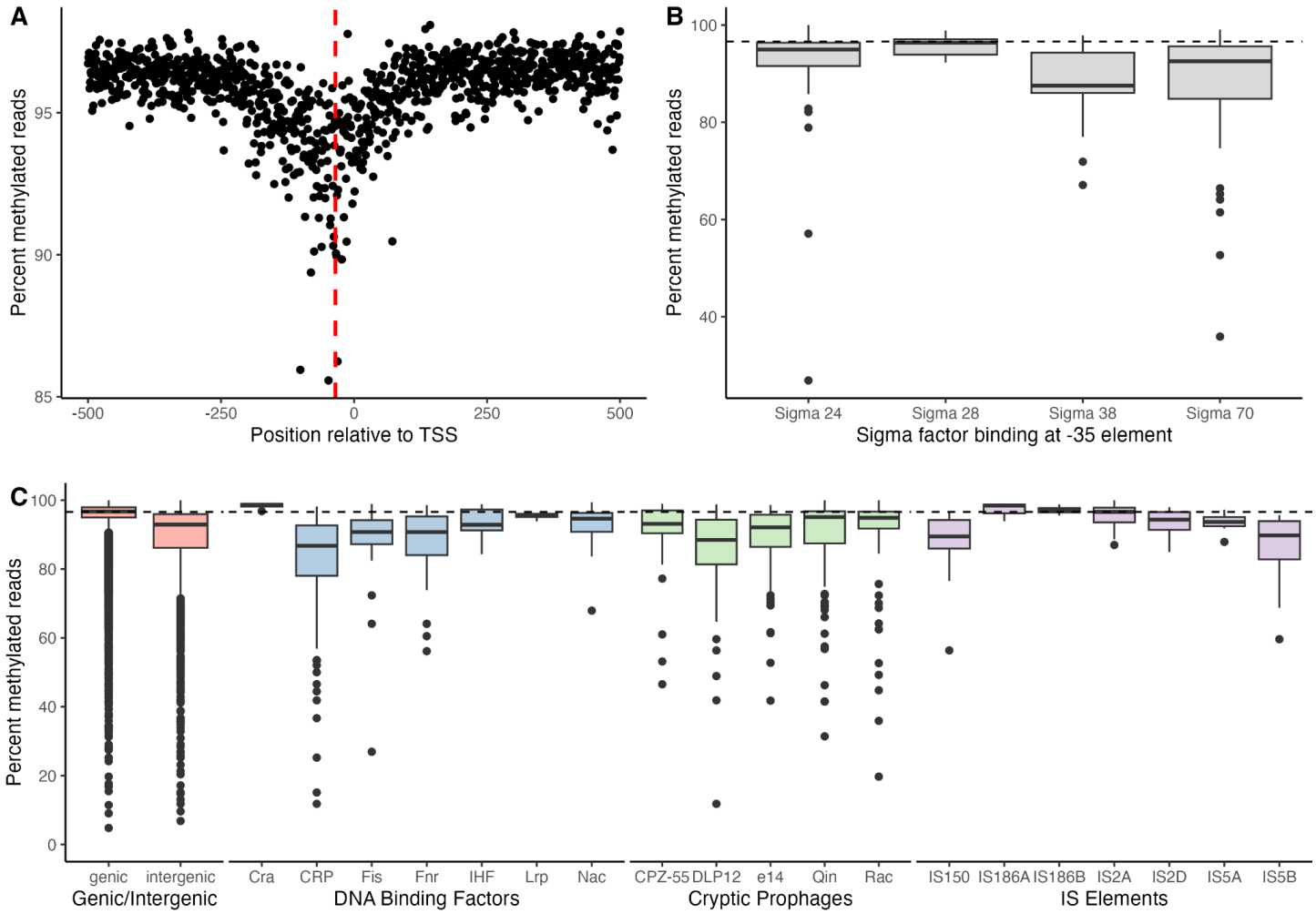

Figure 3. Genome-averaged methylation decreases around core promoter elements. (A) Median methylation at GATC sites aligned relative to the transcription start sites (TSS) for all promoters. Points represent the median methylation for all GATC sites at that position, and the smoothed median (blue line) was calculated with additive quantile regression smoothing. The -35 position relative to the TSS is shown for reference (dashed red line). (B) Comparison of percent methylated reads at -35 elements by specific sigma factor binding. Boxplots show the median, first quartile, and third quartile. Sigma factors were compared to the genome-wide median percent methylation (97%, dashed line) with a one-sample Wilcoxon signed rank test with Benjamini-Hochberg correction. P-values are shown in B and C as *** < 0.001 < ** < 0.01 < * < 0.05. (C) Different genomic features showed variable methylation status. Percent methylation of GATC sites within each feature type was compared to the genome-wide median percent methylation (97%, dashed line) with one-sample Wilcoxon signed rank tests with Benjamini-Hochberg correction.

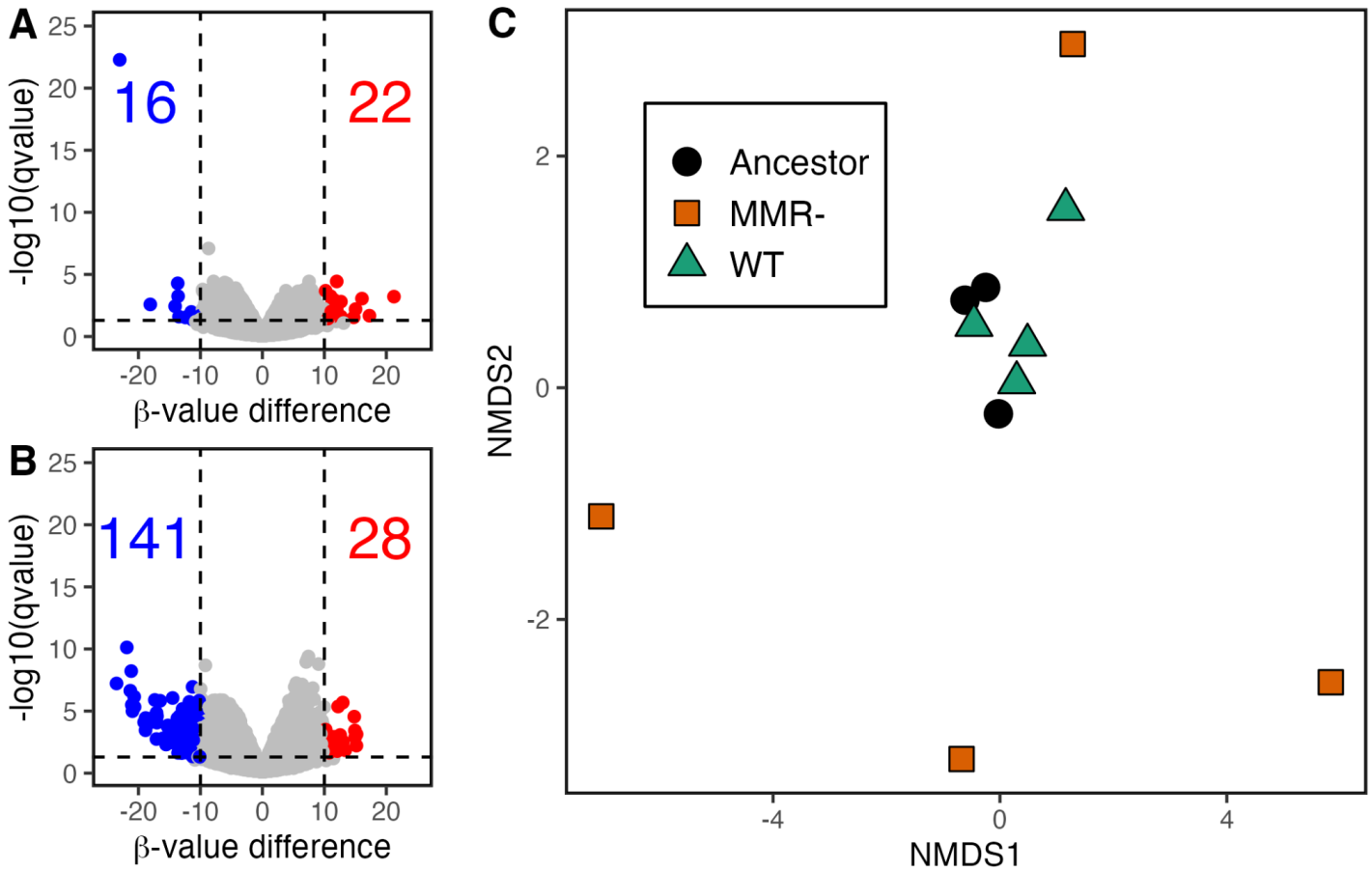

Figure 4. Differentially methylated GATC sites between the ancestor and experimentally evolved (A) WT clones or (B) MMR- clones. Differential methylation was called for WT and MMR- clones compared to the ancestor strain using methylKit. Differential methylation was determined for each site as a percent methylation difference compared to the ancestor greater than 10% and a q-value less than 0.05. (C) Comparison of methylation across all sites between the ancestor and experimentally evolved clones. Euclidean distance between clones was calculated using normalized β-values and ordination was performed with NMDS.

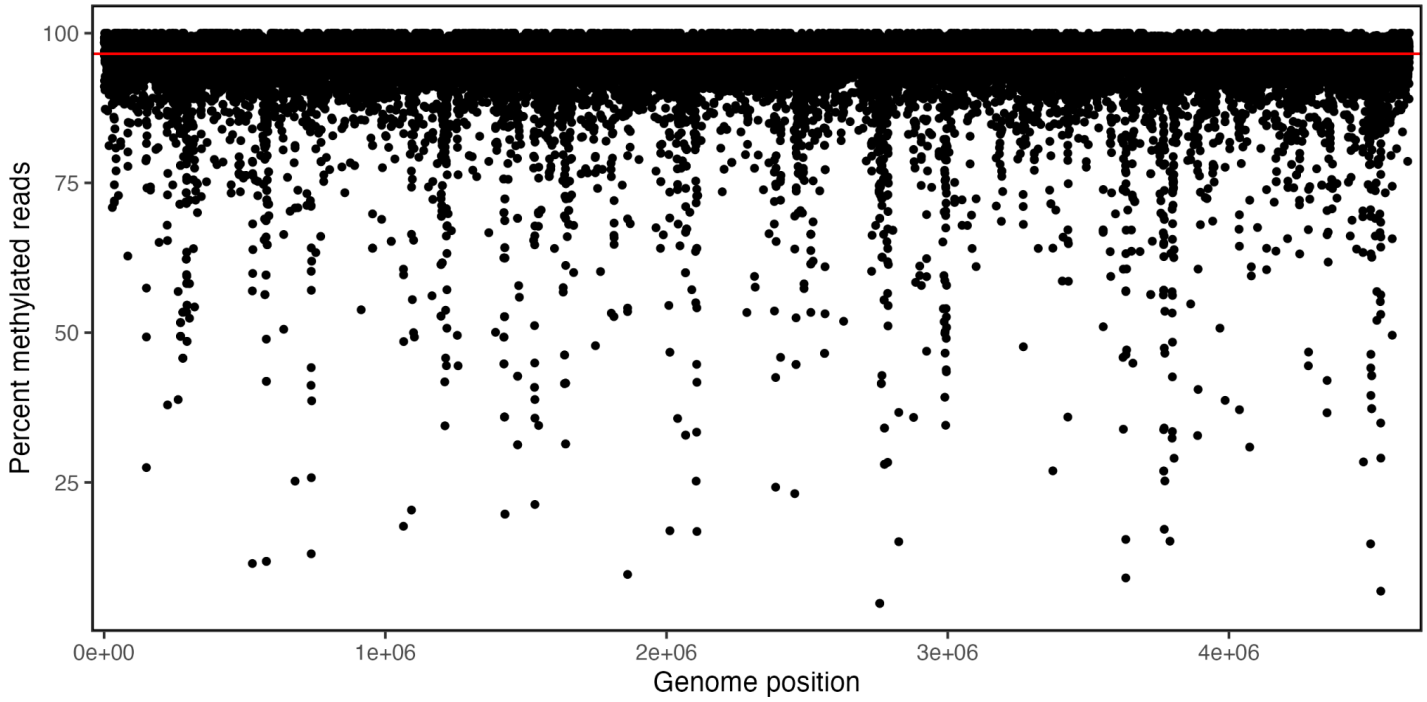

Figure S1. 6mA methylation across the MG1655 genome. Each point represents the mean percent methylation of three sequencing replicates of the K-12 MG1655 ancestor strain at one GATC site. The genome-wide median methylation of 96.6% is shown for reference (red line).

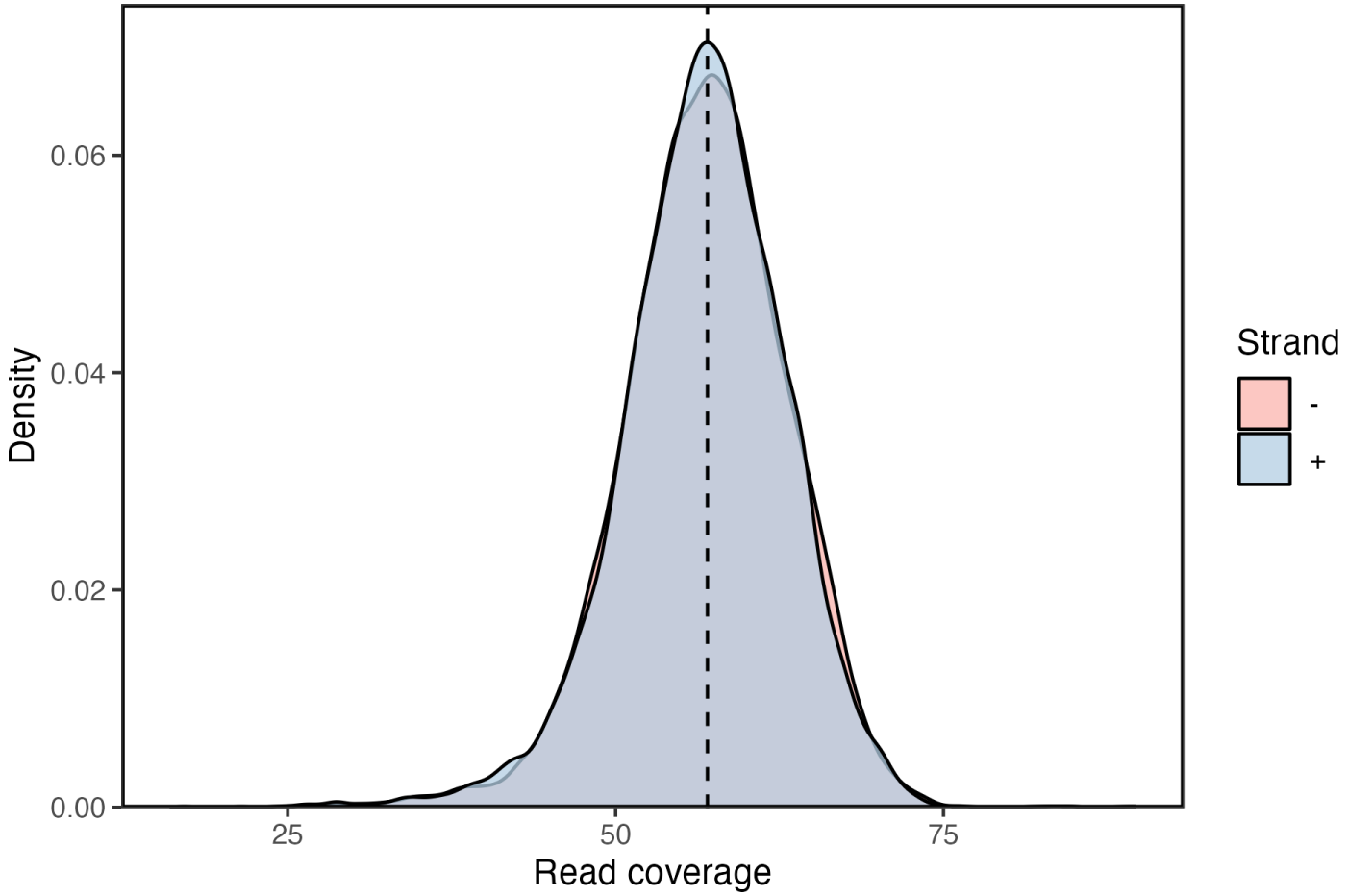

Figure S2. Sequencing coverage of GATC sites by Nanopore sequencing. Coverage distributions for each strand are shown separately, and both distributions have the same median of 57 reads (dashed line).

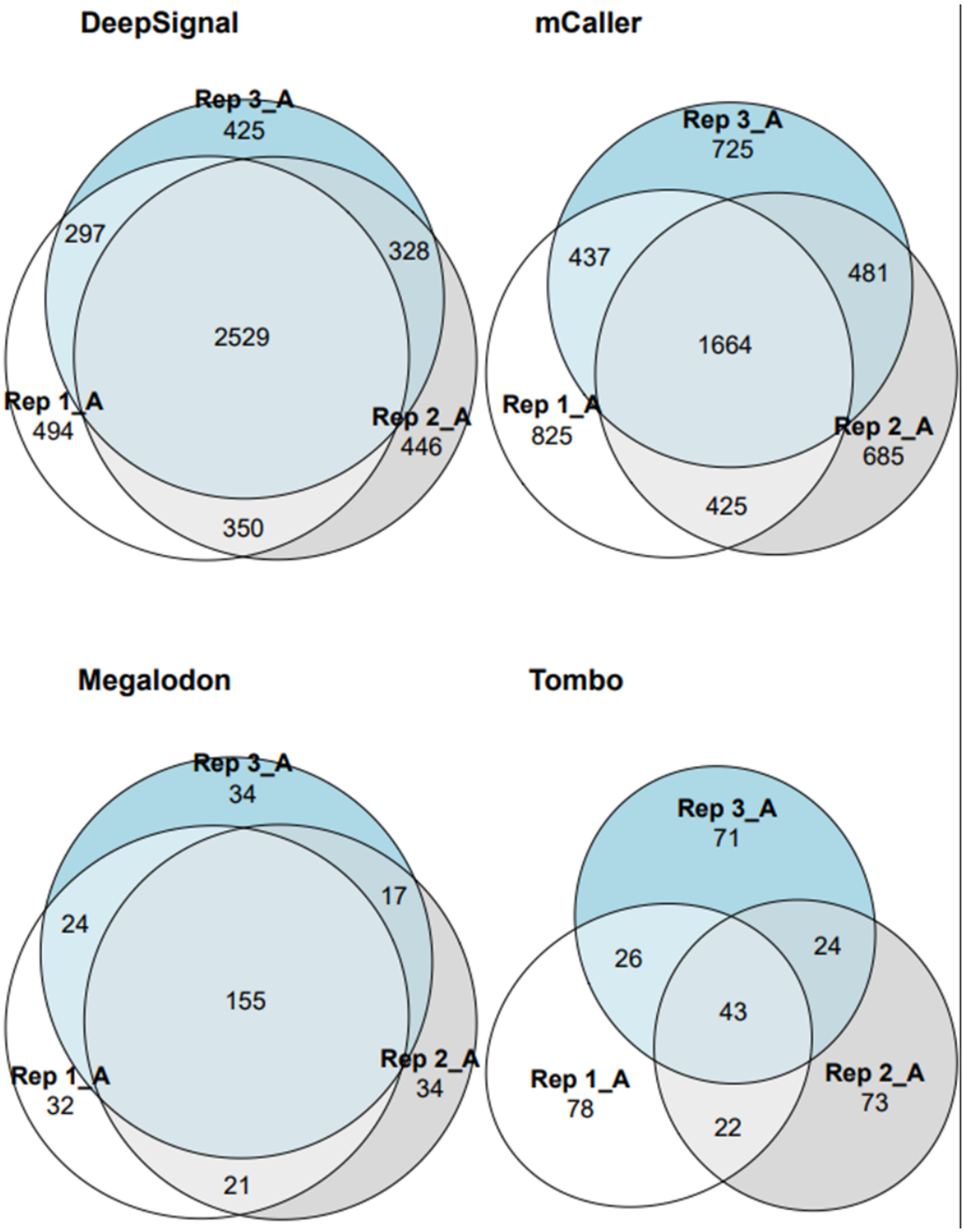

Figure S3. Overlap in sites identified as hypomethylated between three sequencing replicates for each caller. Each area is labeled with the number of GATC sites.

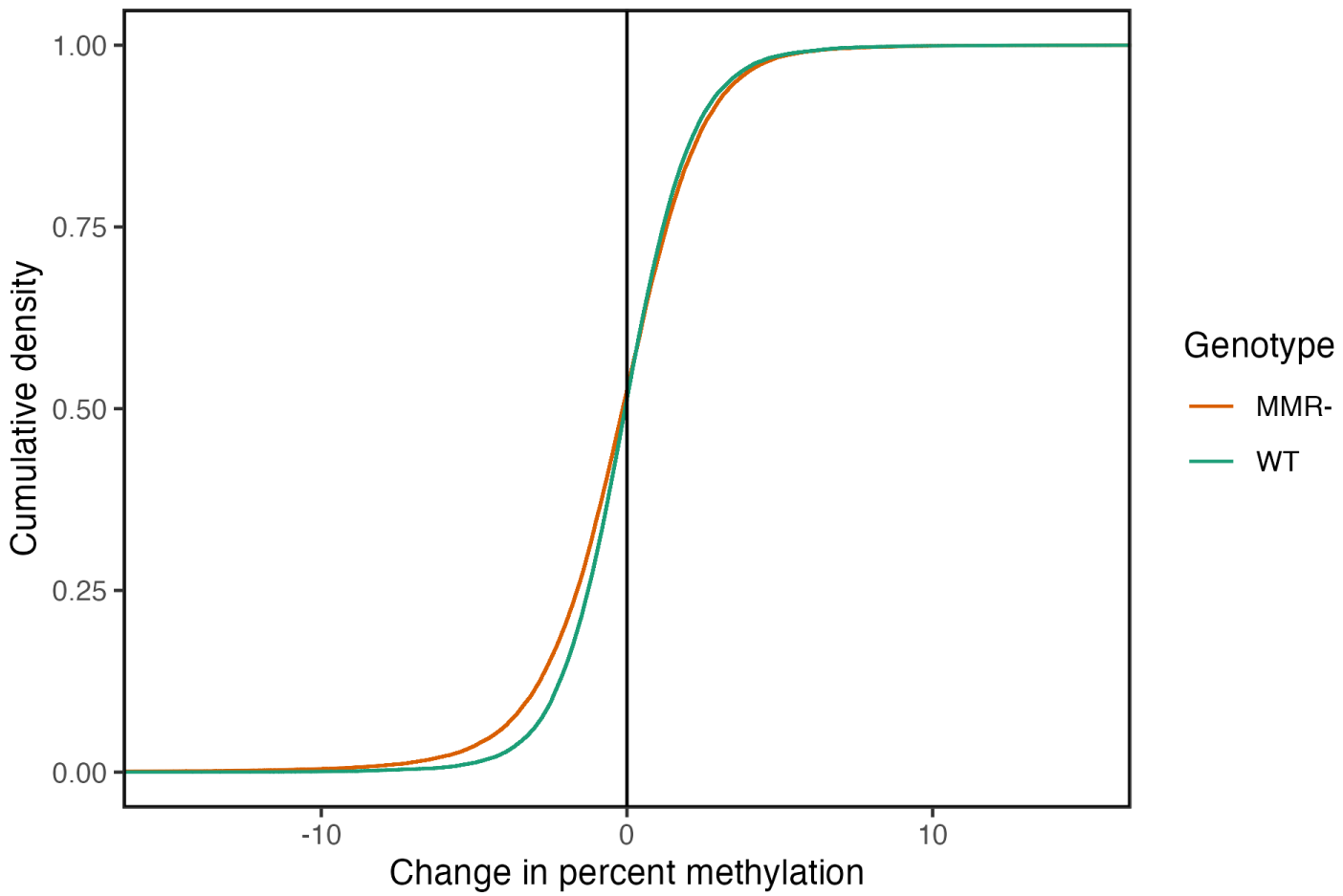

Figure S4. Distribution of changes in per-site methylation between evolved genotypes and their isogenic ancestor. For each GATC site, the difference between percent methylation in the evolved genotype and in the ancestor was calculated, and values were used to calculate empirical cumulative density functions.

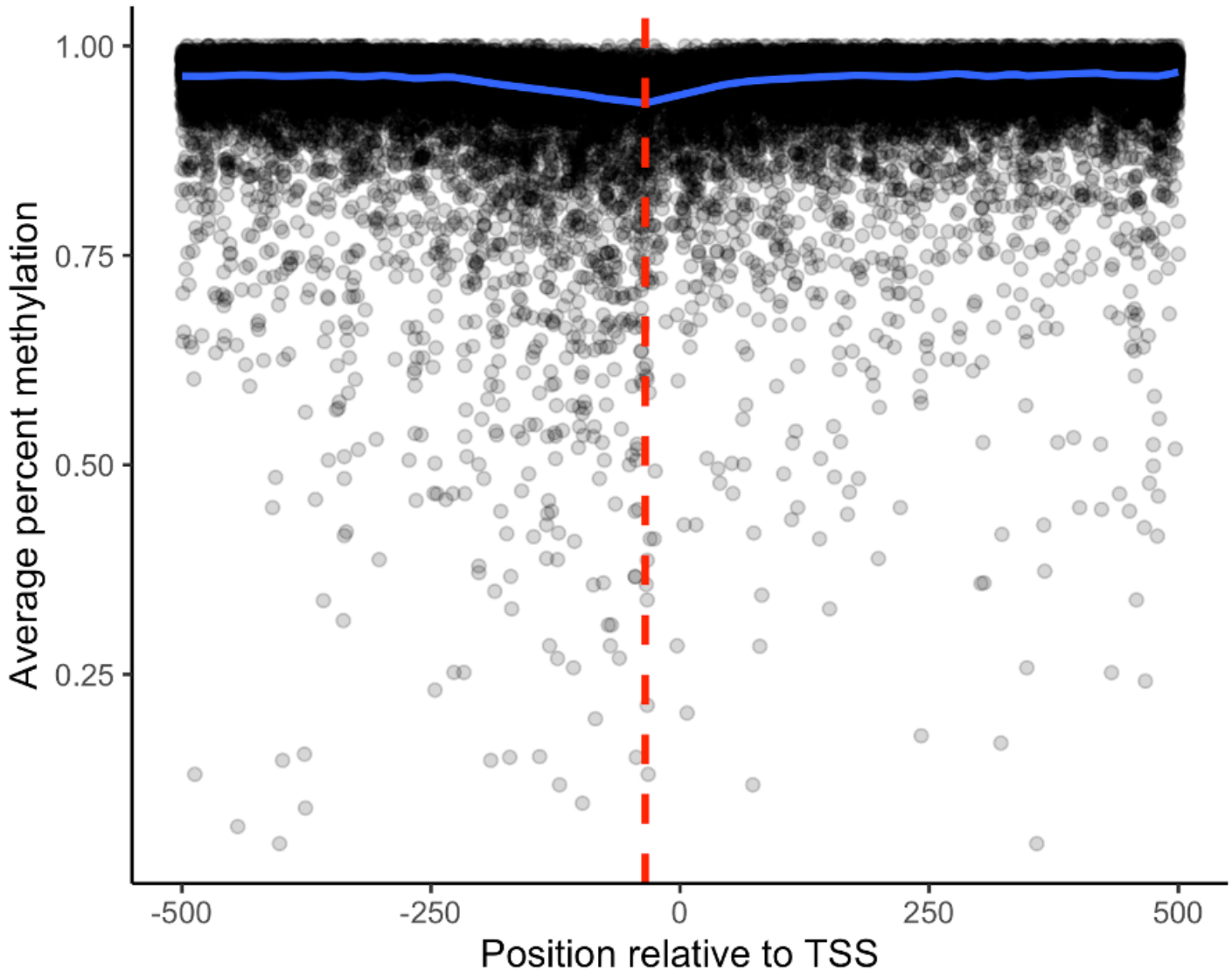

Figure S5. Methylation of all GATC sites relative to the transcription start site (TSS). Each point represents a single GATC site. The smoothed median was calculated by additive quantile regression smoothing (blue line), and the -35 position is shown for reference (dashed red line).

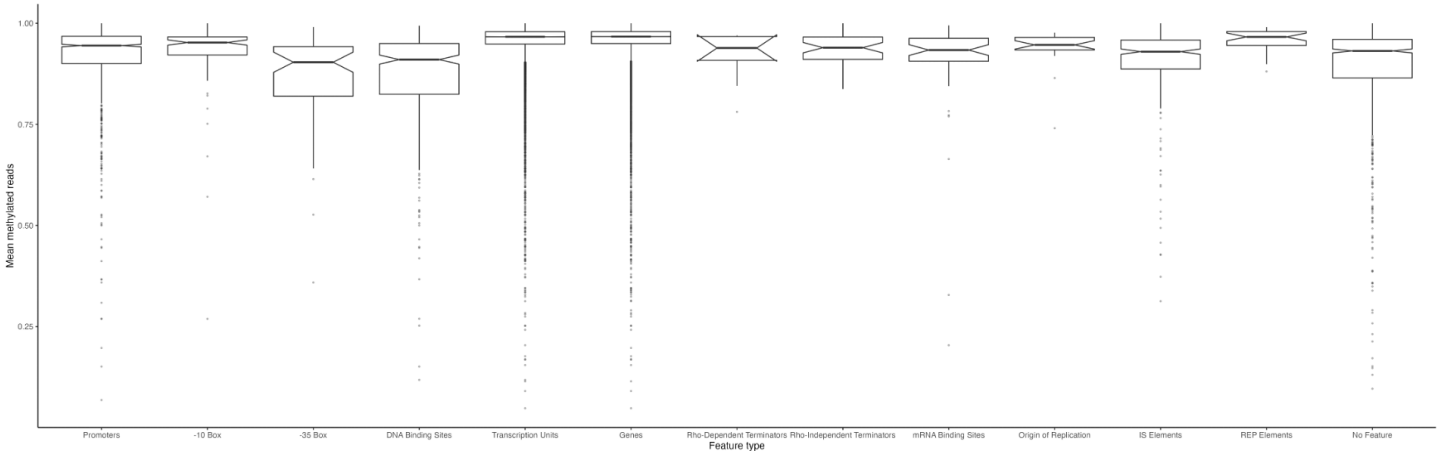

Figure S6. Average methylation at various genomic features. Boxplots show the median, first quartile, and third quartile.

Table 1: Proportion of methylated reads at nonmethylated GATC sites across all callers.

				Proportion of methylated reads			
Location (bp)	**Strand**	**Locus**	**Protein(s) binding at site**	**DeepSignal**	**mCaller**	**Megalodon**	**Tombo**
2825746	+	srlA	srlR/CRP	**0.35**	**0.16**	**0.37**	**0.12**
2825747	-	srlA	srlR/CRP	0.70	**0.18**	**0.15**	**0.38**
3771971	+	mtlA	CRP	**0.30**	**0.44**	**0.25**	0.80
3771972	-	mtlA	CRP	0.62	**0.24**	**0.47**	**0.31**

Table 2. Overrepresentation of differentially methylated adenines in protein-DNA interaction motifs.

		Evolved WT				Evolved MMR-			
Feature Type	# GATC sites in ancestor	Increased methylation	p-value	Decreased methylation	p-value	Increased methylation	p-value	Decreased methylation	p-value
Transcriptional regulator binding sites	313	7	0.095	**12***	**< 0.001**	**11^**	**0.014**	**58***	**< 0.001**
Origin of replication	22	0	1.000	0	1.000	0	1.000	1	0.647
-10 element	116	2	0.465	2	0.428	4	0.187	6	0.299
-35 element	60	1	0.590	1	0.572	1	0.694	**12***	**< 0.001**
Sigma 24 binding sites	67	2	0.294	1	0.590	2	0.380	4	0.299
Sigma 28 binding sites	11	0	1.000	0	1.000	0	1.000	0	1.000
Sigma 32 binding sites	2	0	1.000	0	1.000	0	1.000	0	1.000
Sigma 38 binding sites	17	0	1.000	0	1.000	0	1.000	**3***	**0.043**
Sigma 70 binding sites	86	1	0.694	2	0.313	3	0.254	**12***	**< 0.001**
Genes	35595	312	1.000	275	1.000	442	1.000	952	1.000
Rho-dependent terminators	8	0	1.000	0	1.000	0	1.000	1	0.380
Rho-independent terminators	46	1	0.514	**4***	**0.003**	0	1.000	**6***	**0.011**
IS elements	283	**10^**	**0.002**	**10***	**0.001**	**11^**	**0.007**	**18***	**0.014**
Cryptic prophages	795	**15^**	**0.038**	11	0.249	**24^**	**0.001**	**62***	**< 0.001**
No Feature	1191	**35^**	**< 0.001**	**30***	**< 0.001**	**44^**	**< 0.001**	**134***	**< 0.001**

P-values were calculated by one-tailed Fisher's exact test followed by BH correction. ^ = significant number of sites with increased methylation. * = significant number of sites with decreased methylation.

Table 3. Number of GATC mutations in evolved clones.

Background	Population*	Clones^	Mutations	GATC Gain/Loss	Percent GATC Mutated (This Study)	Percent GATC Mutated (Lee et al, 2012)
MMR	113		257	13	0.051	
		113-1	121	3	0.025	
		113-2	96	3	0.031	
	126		46	1	0.022	
		126-2	313	11	0.035	
		126-4	126	3	0.024	
	Total		959	34	0.035 (0.028 - 0.040)	0.035 (0.028 - 0.043)
WT	125		4	0	0	
		125-6	1	0	0	
		125-7	7	0	0	
	129		4	0	0	
		129-1	8	0	0	
		129-3	9	0	0	
	Total		33	0	0 (0.0 - 0.0)	0.043 (0.025 - 0.067)

*Mutations shared between clones and occurred before subpopulation divergence

^Mutations unique to clones and occurred after subpopulation divergence