Supplementary Tables

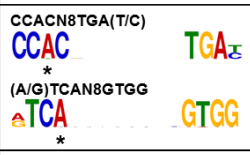
Supplementary Table 1. Coverage statistics for sequencing data on the *Escherichia coli* isolates analyzed in this study.

Strain	Serotype	Location of Isolate	Source of Isolate	# of mapped reads	Mean depth of coverage
55989	O104:H4	Central African Republic	Stool sample from adult with diarrhea	162,912	63.4
C227-11	O104:H4	Denmark	Diarrhea in adult	475,926	190.7
17-2	O3:H2	Chile	Stool sample from child with diarrhea	148,330	52.5
C35-10	O104:H4	Africa	Stool sample from child with diarrhea	223,253	73.3
C682-09	O104:H4	Africa	Stool sample from child with diarrhea	195,513	70.2
C734-09	O104:H4	Africa	Stool sample from child without diarrhea	154,014	59.6
C754-09	O104:H4	Africa	Stool sample from child without diarrhea	80,917	31.5
C760-09	O104:H4	Africa	Stool sample from child without diarrhea	254,893	96.7
C777-09	O104:H4	Africa	Stool sample from child with diarrhea	74,932	31.1

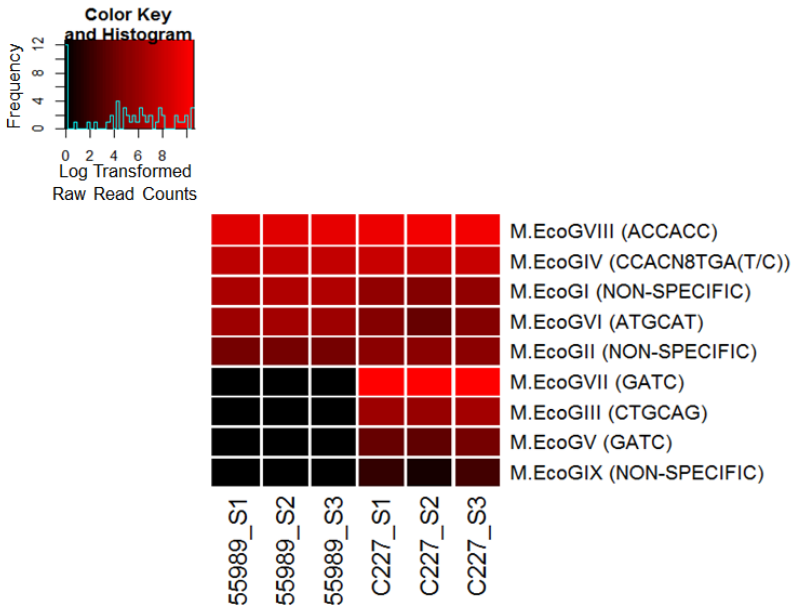
Supplementary Table 2. Summary of the number of base modification detections by strand and motif made in the C227-11 SMRT sequencing data.

Base-level detections	Number of sites tested	Number of sites detected at FDR<0.01	Percentage of sites detected at FDR<0.01
A	2,579,260	49,311	1.91%
C	2,649,614	1,407	0.053%
G	2,646,709	833	0.031%
T	2,577,967	421	0.016%
Motif-level detections	**Detected in both strand**	**Detected in only one strand**	**Detected in neither strand**
5'—GATC—3'	18,647 (89.2%)	2,030 (9.7%)	220 (1.1%)
5'—CTGCAG—3'	1,157 (93.1%)	77 (6.2%)	9 (0.7%)
5'—CCACN8TGA(T/C)—3'	439 (97.6%)	11 (2.4%)	0 (0.0%)
5'—ACCACC—3'	N/A	4130 (100%)	N/A
5'—CCWGG—3'	0 (0%)	92 (0.7%)	13351 (99.3%)

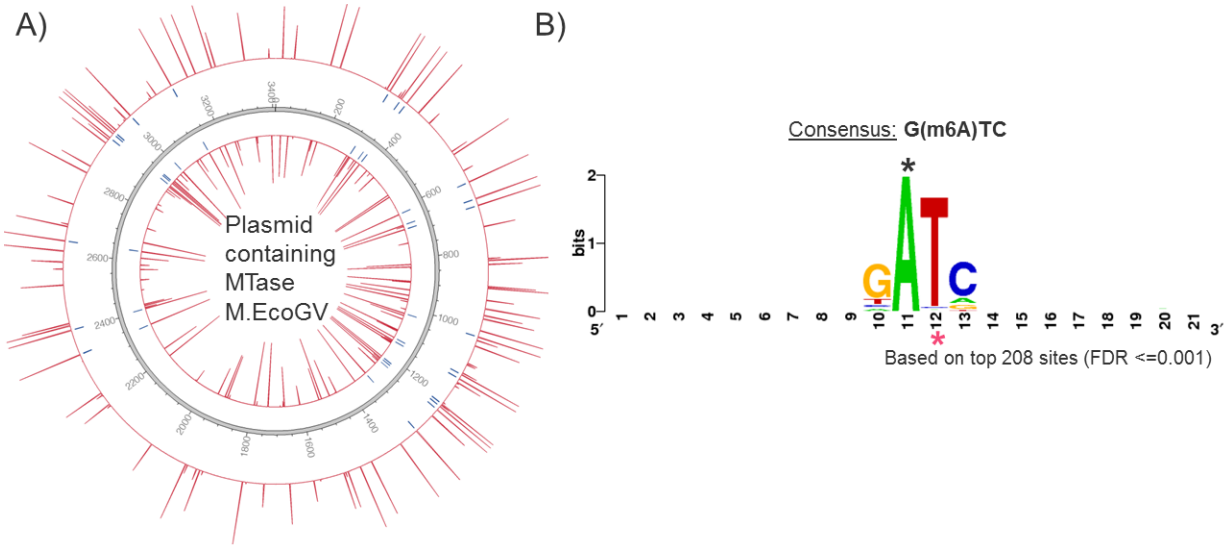
Supplementary Figures

Methyltransferase	Motif (weblogo)	C227 outbreak	55989	682-09	17-2	734-09	760-09	35-10	042	1010
M.EcoGI	**Non-specific**	■	■	■	■	■	■	■	■	■
M.EcoGII	**Non-specific**	■	■	■	■	■	■	■	■	■
M.EcoGV		■		■	■	■		■		
M.EcoGVIII		■	■	■	■	■	■	■	■	
M.EcoGVI		■	■	■	■	■	■	■	■	■
M.EcoGIV		■	■	■		■	■	■		
M.EcoGIII		■								
M.EcoGVII		■								
M.EcoGIX	**Non-specific**	■		■		■	■	■	■	■
M.EcoGDam		■	■	■	■	■	■	■	■	■

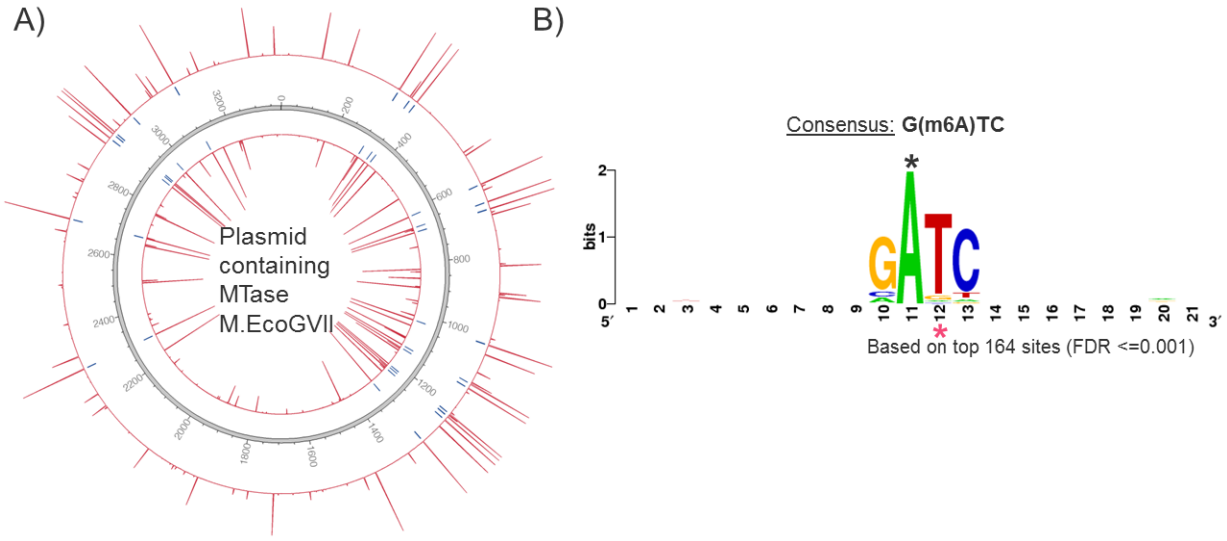
Supplementary Figure 1. Pattern of methyltransferase genes across the nine O104 *E. coli* strains sequenced. The first column indicates the methyltransferase gene name in strain C227, the second column indicates the motif identified in the C227 outbreak strain that was associated with the corresponding methyltransferase as determined by the expression of that gene in the plasmid expression system (black asterisks indicate modified A residues on the sense strand), and the remaining columns indicate the presence (blue fill) or absence (white fill) of the corresponding methyltransferase in the indicated O104 strain. The bolded methyltransferase M.EcoGVI was associated with the ATGCAT motif in the plasmid expression system, but despite being present in all O104 strains sequenced, this enzyme was not observed as active in any of the strains. Similarly, the non-specific A-MTases, M.EcoGI, M.EcoGII, and M.EcoGIX, were also not expressed in any of the strains.



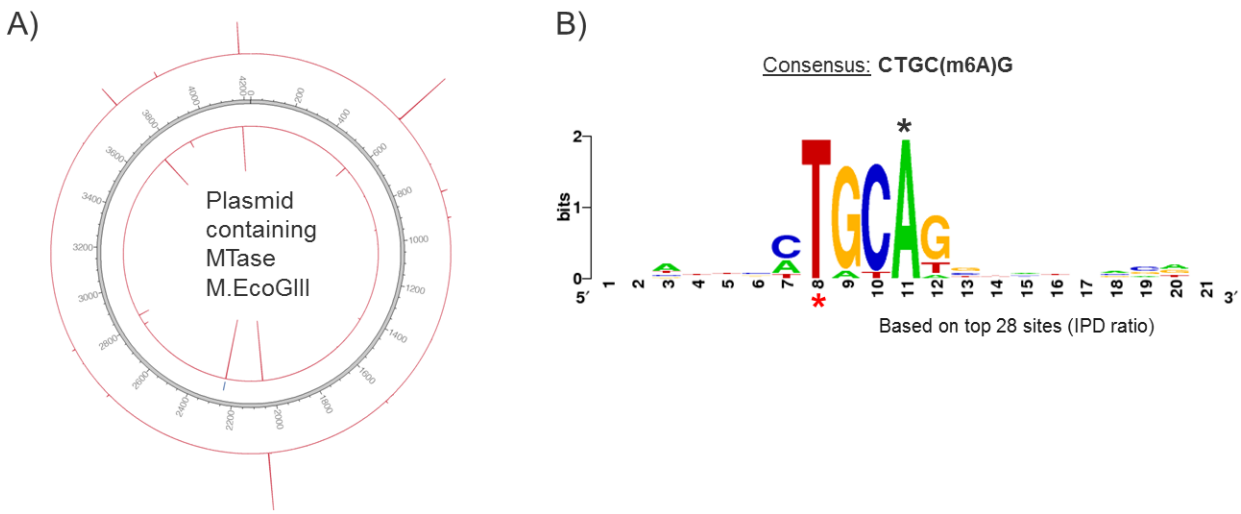
Supplementary Figure 2. Expression levels of the 9 predicted MTases detailed in Table 1 of the main text as measured in three samples each of the C227 and 55989 strains of *E. coli*. RNA from each sample was isolated, sequenced and analyzed to determine transcript abundances for all genes of interest. RNA sequencing counts for each ORF determined in each sample were log transformed (natural log). The upper left panel is the color key and histogram for the expression values, with black indicating no expression detected and bright red indicating high levels of expression (the x-axis represents the normalized, log transformed raw read counts, and the y-axis represents the number of elements in the expression matrix, right panel, with the indicated read counts). The genes M.EcoGIII, M.EcoGV, M.EcoGVII, and M.EcoGIX are present in C227 but not in 55989, consistent with the RNA sequencing data in which no reads for any of these four genes detected in strain 55989.



Supplementary Figure 3. Validation of M.EcoGV identified in the German outbreak strain as the methyltransferase M.EcoK*dam-like* specifically targeting the GATC context. **A)** Plasmid containing M.EcoGV depicted as a circos plot, with the inside of the annulus representing the coordinates of the plasmid, the blue hash marks indicating the A residues in a GATC context, and the two red curves representing the –log10(p value) for the likelihood ratio model for the two DNA strands. The A residues in all 42 of the GATC contexts in the plasmid were detected as modified at FDR ≤ 0.001. **B)** Weblogo plot based on the top 208 A residue sites detected as modified (FDR ≤ 0.001), with GATC representing the dominate motif, demonstrating the specificity of M.EcoK*dam* for this motif.

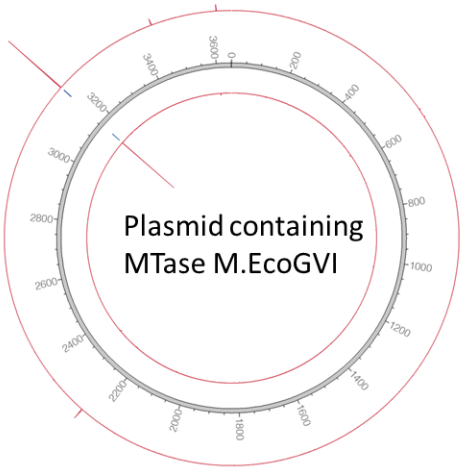


Supplementary Figure 4. Validation of M.EcoGVII as the methyltransferase M.EcoK*dam-like* specifically targeting the GATC context. **A)** Plasmid containing M.EcoGVII depicted as a circos plot, with the inside of the annulus representing the coordinates of the plasmid, the blue hash marks indicating the A residues in a GATC context, and the two red curves representing the –log10(p value) for the likelihood ratio model for the two DNA strands. The A residues in all 40 of the GATC contexts in the plasmid were detected as modified at FDR ≤ 0.001. **B)** Weblogo plot based on the top 208 A residue sites detected as modified (FDR ≤ 0.001), with GATC representing the dominate motif, demonstrating the specificity of M.EcoK*dam* for this motif.

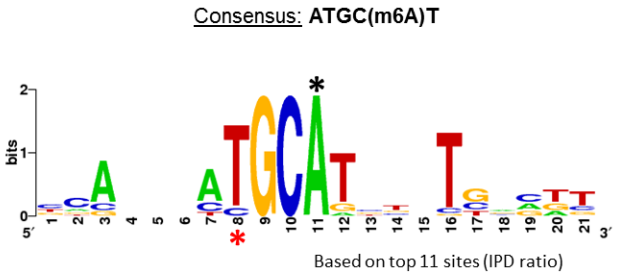


Supplementary Figure 5. Validation of M.EcoGIII as the MTase targeting the CTGCAG motif. **A)** Plasmid containing M.EcoGIII depicted as a circos plot, with the inside of the annulus representing the coordinates of the plasmid, the blue hash marks indicating the A residues in a CTGCAG context, and the two red curves representing the –log10(p value) for the likelihood ratio model for the two DNA strands. The A residues in all 2 of the CTGCAG contexts in the plasmid were detected as modified at an FDR ≤ 0.001. **B)** Weblogo plot based on the top 28 A residue sites detected as modified (FDR ≤ 0.001), with CTGCAG representing the dominate motif, demonstrating the specificity of M.EcoGIII for this motif.

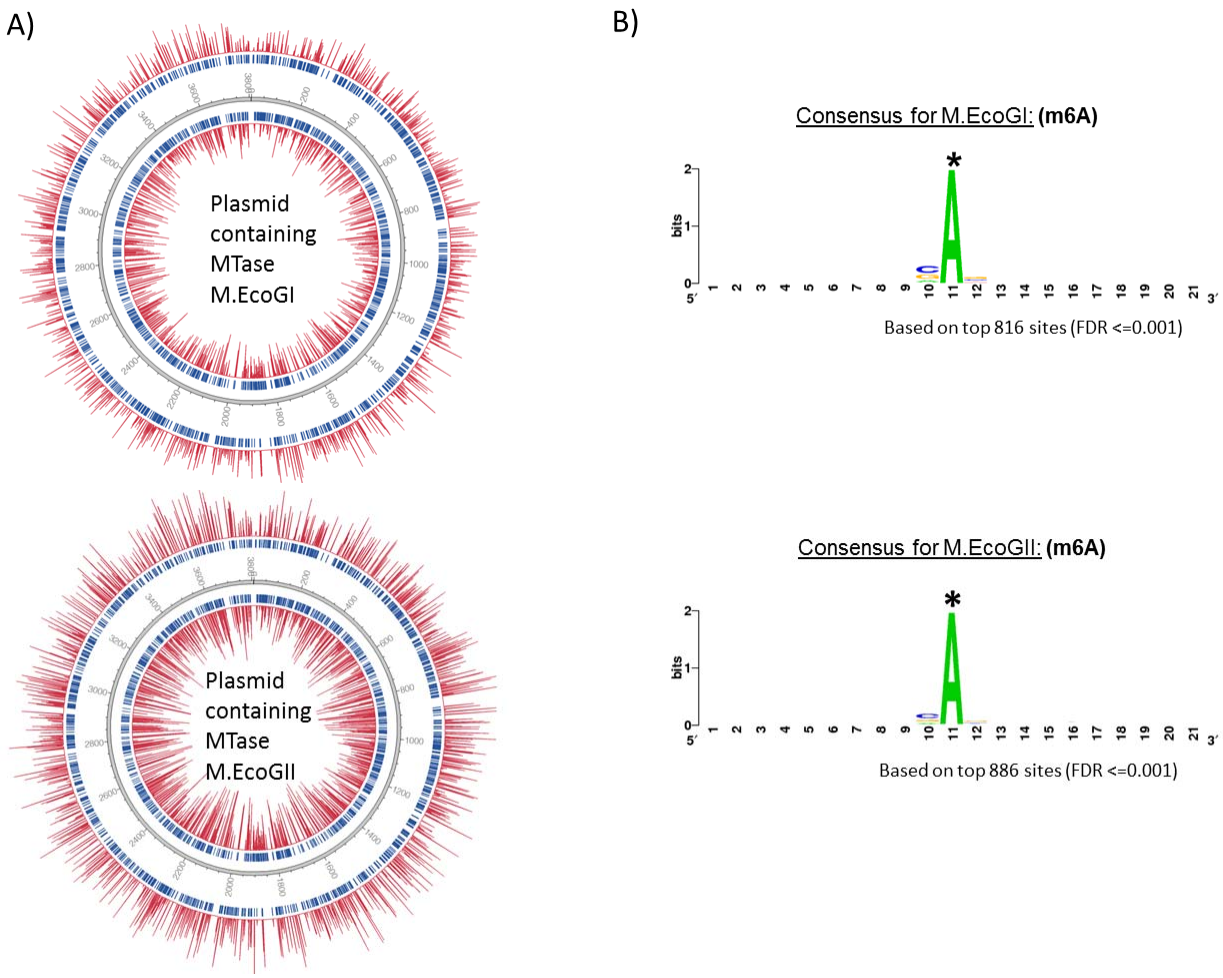
A)



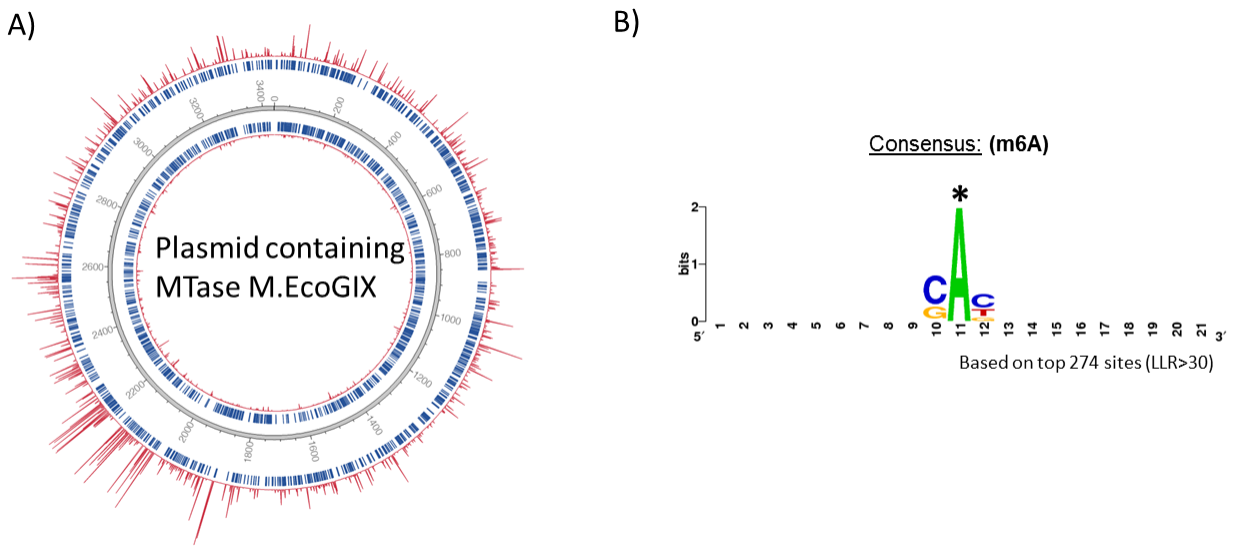
B)



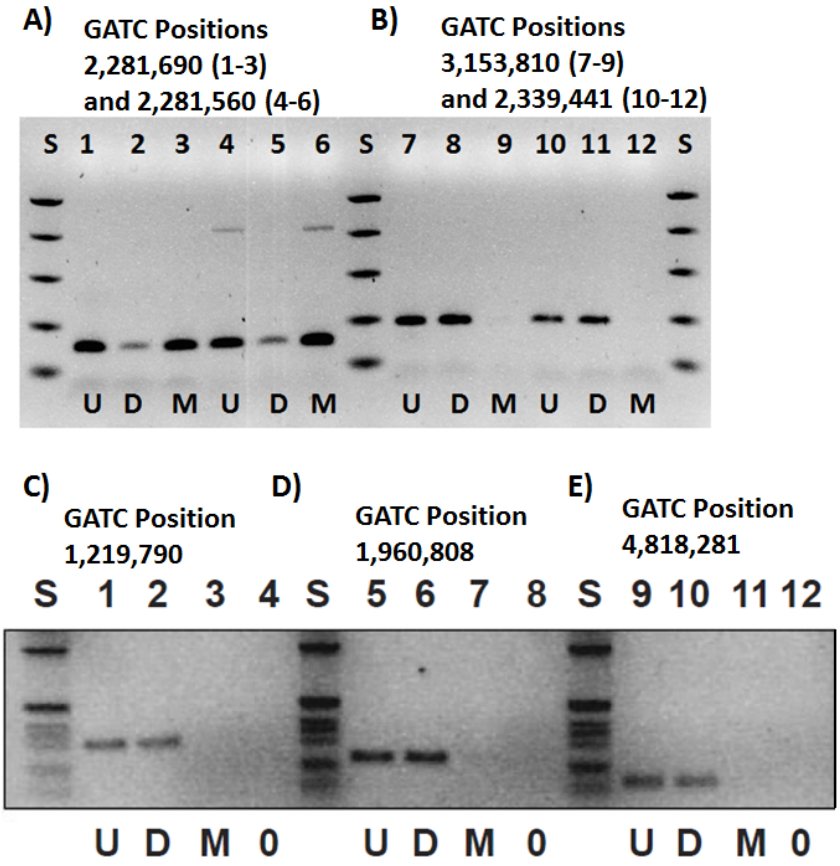
Supplementary Figure 6. Validation of M.EcoGVI as an MTase targeting the ATGCAT motif. **A)** Plasmid containing M.EcoGVI depicted as a circos plot, with the inside of the annulus representing the coordinates of the plasmid, the blue hash marks indicating the A residues in a ATGCAT context, and the two red curves representing the –log10(p value) for the likelihood ratio model for the two DNA strands. Both the ATGCAT contexts in the plasmid were detected as modified at an FDR ≤ 0.001. **B)** Weblogo plot based on the top 11 A residue sites detected as modified (FDR ≤ 0.001), with ATGCAT representing the dominate motif, demonstrating the specificity of M.EcoGVI for this motif.



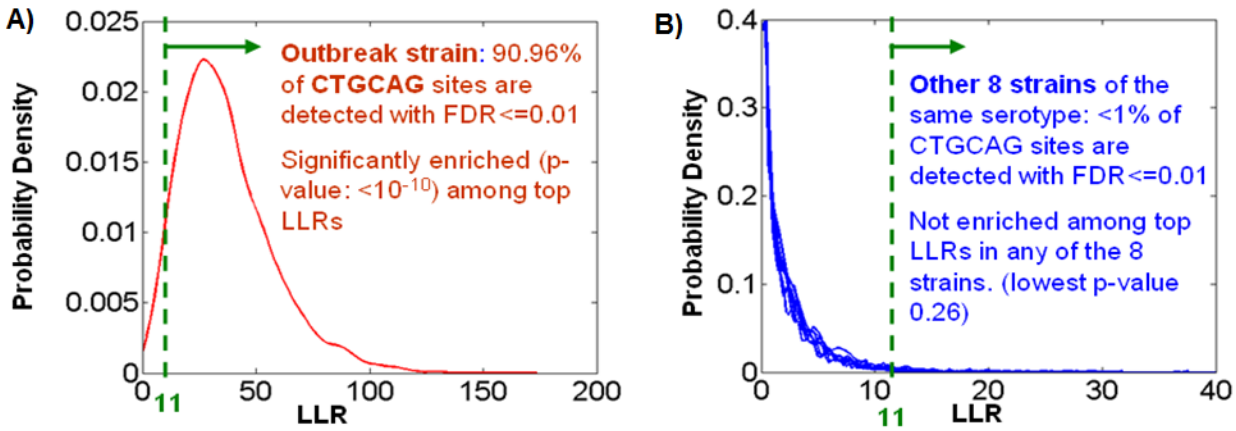
Supplementary Figure 7. Validation of M. Eco GI and M.EcoGII as non-specific MTases non-discriminantly targeting A residues. **A)** Plasmids containing M.EcoGI and M.EcoGII depicted as circos plots, with the inside of the annulus representing the coordinates of the plasmid and the two red curves representing the –log10(p value) for the likelihood ratio model for the two DNA strands. Of the 1,869 A bases in the M.EcoGI plasmid, 1,500 (80.26%) were detected as modified at an FDR < 0.001. Of the 1,870 A bases in the M.EcoGII plasmid, 1,648 (88.13%) were detected as modified at an FDR ≤ 0.001. **B)** Weblogo plots based on the top 816 A base sites for M.EcoGI and 886 A base sites for M.EcoGII detected as modified, with A representing the dominate motif, demonstrating the non-specificity of these MTases.



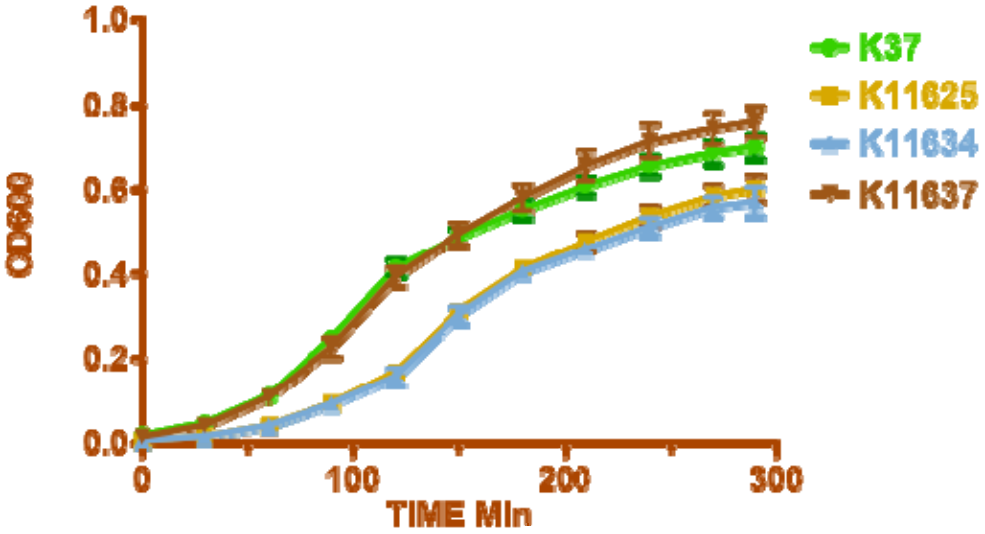
Supplementary Figure 8. Validation of M.EcoGIX as a non-specific MTase non-discriminantly targeting A residues on a single strand. **A)** Plasmid containing M.EcoGIX depicted as a circos plot, with the inside of the annulus representing the coordinates of the plasmid and the two red curves representing the –log10(p value) for the likelihood ratio model for the two DNA strands. Signal KV was only detected on the plus strand of the plasmid, suggesting that this enzyme is strand specific. Of the 1659 A residues on the plus strand of the plasmid, 274 (16.5.%) were detected as modified at an FDR ≤ 0.001. Of these 274 sites, 237 (86.5%) were from the positive strand, suggesting a strand-specific bias (sign test p < 10e-52) of the methylase. **B)** Weblogo plot based on the top 274 A residue sites detected as modified (FDR ≤ 0.001), with no sequence motifs identified as enriched in the plasmid sequence, demonstrating the non-specificity of M.EcoGIX.



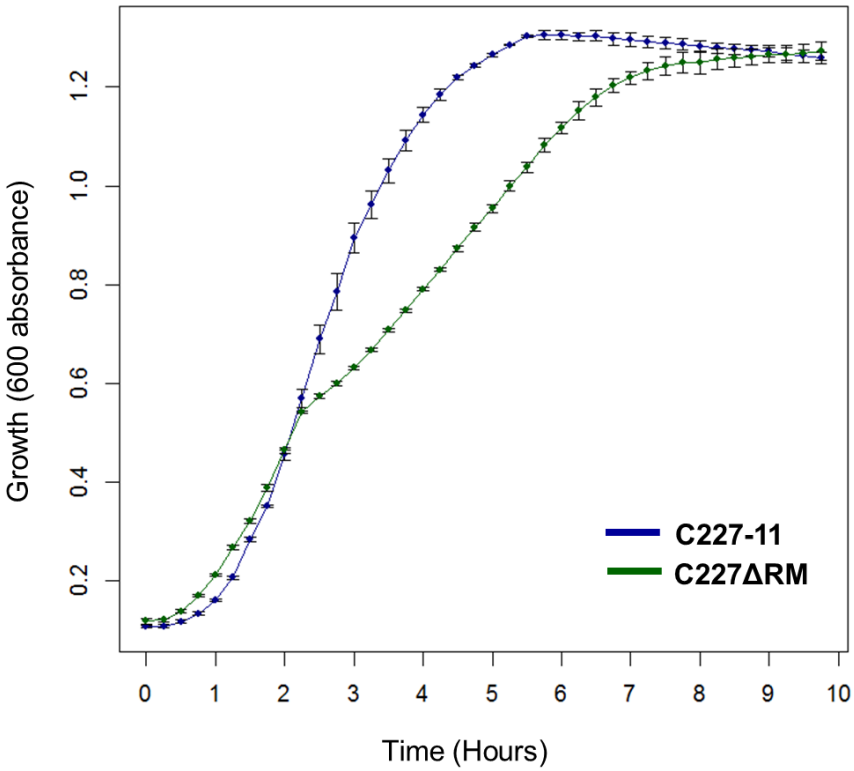
Supplementary Figure 9. Assessment of *dam* methylation status of selected GATC sequences within the *E. coli* C277 genome. Samples of C227 genomic DNA (100ng) were incubated in the absence (U) or presence of 25 units of DpnI (D) or MboI (M) restriction endonucleases at 37°C for 60min (panels A and B) and 90 minutes (panels C-E) (in 0.1 ml of 1 x NEB Buffer 4). After heat inactivation of endonucleases at 80°C (20 min) individual loci containing GATC sequences were PCR-amplified (25 cycles) using Phusion-HF DNA polymerase, locus-specific oligonucleotide primers and 1 µl of sample as template. Cycling parameters were 25 cycles of 98°C (10 seconds), 65°C (10 seconds) and 72°C (30 seconds). Amplification products were resolved by electrophoresis using 1.5% (w/v) agarose in 1 x TAE buffer. Size standards (S) are 766, 500, 300, 150 and 50 bp, respectively, for panels A and B, and 1,018, doublet at 506/517, 396, 344, 298, 220, 201, 154, 134, and 75bp, respectively, for panels C-E. Samples in panel A correspond to GATC sequences that were shown by SMRT sequencing to be heavily methylated, which should render them sensitive to DpnI restriction and resistant to MboI restriction, respectively (amplicon size = 101 bp). The faint band in D column for panel A reflects the presence of 6mA but also indicates that not all molecules at these sites are modified, consistent with the mixture predictions at GATC sites described in the main test. The remaining panels B-E correspond to GATC sequences at the indicated positions that were shown by SMRT sequencing to be unmethylated, which renders them resistant to DpnI restriction and sensitive to MboI restriction, respectively (amplicon sizes = 149, 300, 241, and 162bp for panels B-E).



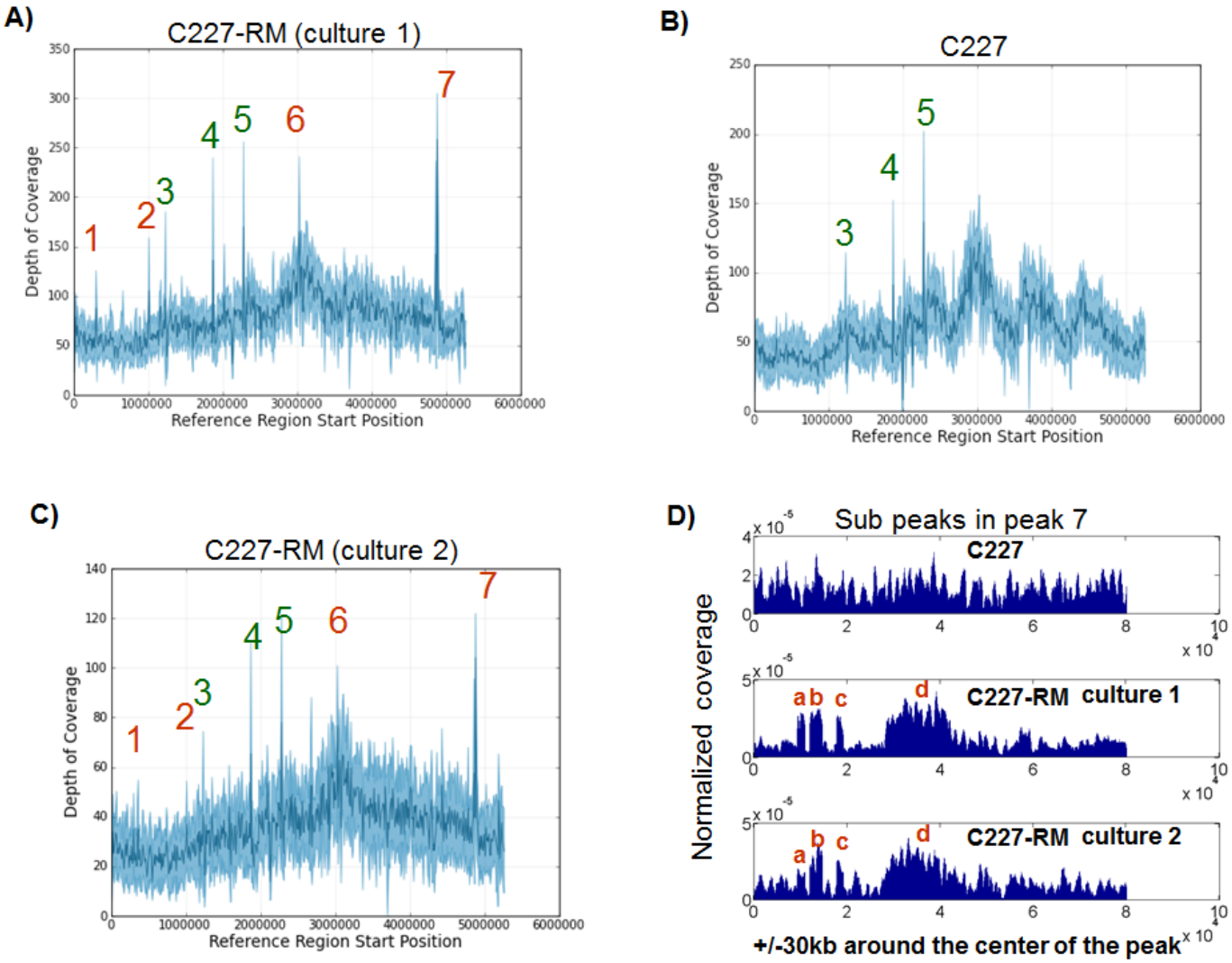
Supplementary Figure 10. The detection distribution of A residues in the CTGCAG motif in the outbreak strain (a) and the other eight strains (b). In each distribution plot the loglikelihood ratio (LLR) used to detect KV events is on the x-axis, and the probability density is on the y-axis. The vertical green dashed line indicates the LLR threshold for a 5% FDR. While > 93% of the CTGCAG motifs were detected in strain C227-11 (the LLR distribution being very biased away from an LLR of zero, the expected LLR value under the null hypothesis of no kinetic variation), < 1% of these sites were detected in the non-outbreak strains (LLR distribution concentrated at zero), reflecting no increased enrichment of CTGCAG sites beyond what would be expected by chance in the case of the non-outbreak strains, and a > 47-fold enrichment in the outbreak strain case.



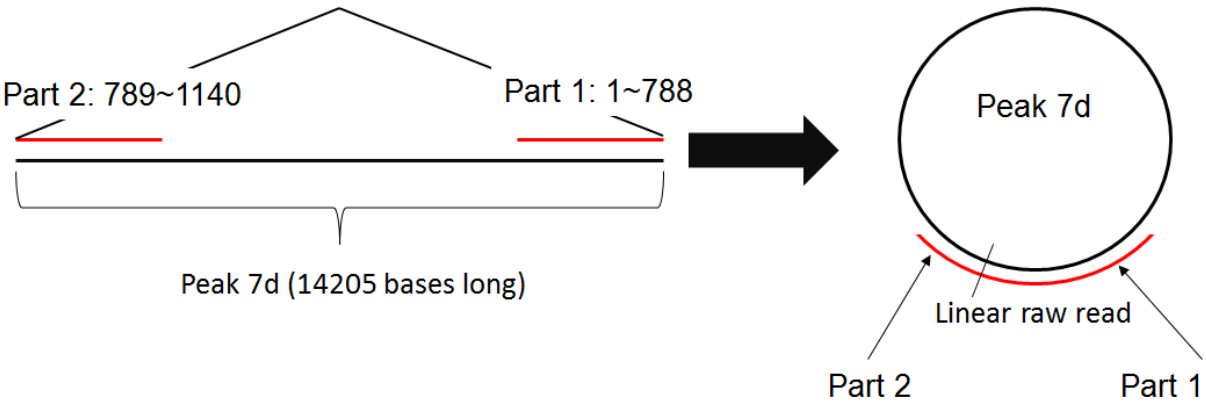
Supplementary Figure 11. The K37(ϕ104X) lysogen has impaired growth compared with its parental strain due to the action of the PstI-like methylase, M.EcoGIII. The growth assay was carried out in LB broth at 37° C and was measured by a Spectramax 250 instrument. The x-axis = time in minutes and the y-axis = optical density (600nm absorbance). Growth curves for 4 different strains are represented: green = K37 (standard K12 laboratory strain), gold = K37(ϕ104X) lysogen (K11625), blue = K37(ϕ104X) lysogen sans Pst1-like endonuclease (K11634), red = K37(ϕ104X) lysogen sans Pst1-like nuclease and methylase (K11637).



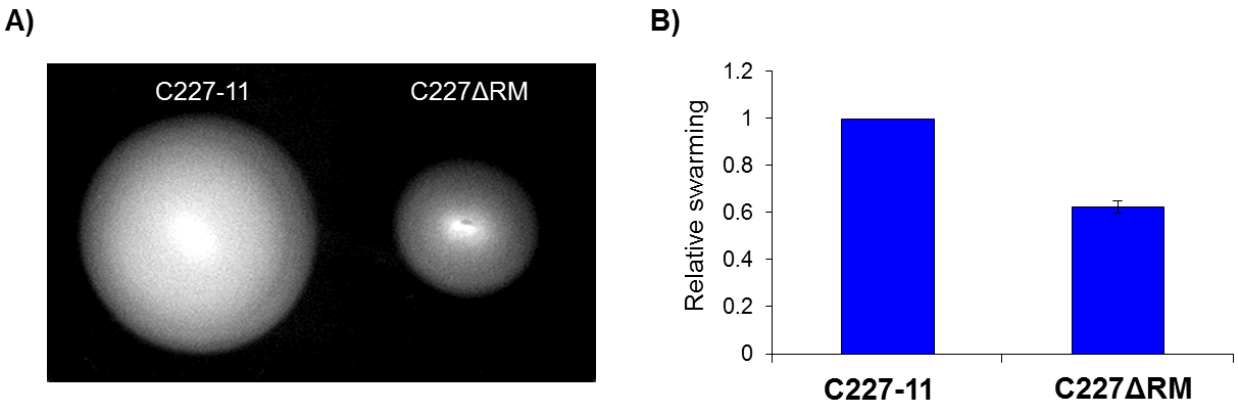
Supplementary Figure 12. Impact of M.EcoGIII on growth phenotypes. Growth curves for wildtype C227-11 (blue curve) and C227ΔRM (green curve) grown in LB medium. The most significant difference in growth occurs at 4.5 hours with a difference of 0.35 ($p = 8.4 \times 10e\text{-}7$).



Supplementary Figure 13. Coverage plots reflecting the mapping of DNA sequencing reads from C227ΔRM (panel A), C227 (panel B) and an independent culture of C227ΔRM (panel C) to a reference of the German *E. coli* outbreak strain. Among the seven high-coverage peaks noted in (A), four peaks (red) are unique to C227ΔRM, while the other three (green) also exist in C227. (D) A higher resolution view of the coverage over the +/-30kb region around peak 7. Four subpeaks are consistently observed in two independent cultures of C227ΔRM but not in C227.



Supplementary Figure 14. Raw DNA reads support the presence of circular constructs (plasmids). A representative read (red segment), out of 6 reads identified supporting circularization, aligns to the start and end of the 7d region depicted in Supplementary Figure 12 (panel D). This read is 1,140 bases in length, but is spanning greater than 14,000 bases through region 7d, suggesting that sub-peak 7d corresponds to an independent circular construct in addition to a linear segment in the C227-11 genome.



Supplementary Figure 15. Swarming motility is impaired in C227ΔRM compared to wildtype C227-11. A) Swarm assay of wild-type and C227ΔRM on a 0.3% agar LB medium showing that the C227ΔRM strain is partially impaired in swarming ability compared to wildtype. B) Swarming distance of wildtype and C227ΔRM strains on 0.3% agar LB medium swarming plates. Data is an average of 14 individual swarming assays; for each assay, swarming by the mutant was calculated relative to that of the wt, which was set to 1. Error bar shows the standard error of the mean.