nature biotechnology

Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing

Gang Fang[1,11], Diana Munera[2,3,11], David I Friedman[4], Anjali Mandlik[2,3], Michael C Chao[2,3], Onureena Banerjee[5], Zhixing Feng[6–8], Bojan Losic[9], Milind C Mahajan[9], Omar J Jabado[9], Gintaras Deikus[9], Tyson A Clark[5], Khai Luong[5], Iain A Murray[10], Brigid M Davis[2,3], Alona Keren-Paz[9], Andrew Chess[9], Richard J Roberts[10], Jonas Korlach[5], Steve W Turner[5], Vipin Kumar[1], Matthew K Waldor[2,3] & Eric E Schadt[9]

Single-molecule real-time (SMRT) DNA sequencing allows the systematic detection of chemical modifications such as methylation but has not previously been applied on a genome-wide scale. We used this approach to detect 49,311 putative 6-methyladenine (m6A) residues and 1,407 putative 5-methylcytosine (m5C) residues in the genome of a pathogenic *Escherichia coli* strain. We obtained strand-specific information for methylation sites and a quantitative assessment of the frequency of methylation at each modified position. We deduced the sequence motifs recognized by the methyltransferase enzymes present in this strain without prior knowledge of their specificity. Furthermore, we found that deletion of a phage-encoded methyltransferase-endonuclease (restriction-modification; RM) system induced global transcriptional changes and led to gene amplification, suggesting that the role of RM systems extends beyond protecting host genomes from foreign DNA.

The information content of the genetic code is not limited to the primary nucleotide sequence but is also conveyed by chemical modifications of individual bases. These modifications expand the genetic alphabet beyond the canonical deoxyribonucleotides (A, G, C and T) to include molecules such as m5C, 5-hydroxymethylcytosine (5-hmC), N4-methylcytosine (m4C), m6A and N6-carboxymethyladenine[1–7]. In eukaryotes, epigenetic DNA modifications affect diverse cellular and developmental processes[8–10]. DNA modification also has a crucial role in bacterial biology and has been shown to regulate chromosome replication, transcription and virulence, and contributes to population diversity through control of phase variation[11–16]. However, although developments in sequencing technologies continue to increase throughput and reduce costs of sequencing at rates exceeding those predicted by Moore's law, the ability to unambiguously characterize variations in the chemical composition of the nucleotide bases in DNA sequences has lagged behind. Although detection of changes in m5C after samples have been bisulfite-treated is possible using next-generation sequencing technologies[17,18] and assays for the genome-wide detection of 5-hmC have been developed[19,20], other modifications cannot be systematically detected using current sequencing instruments[21].

Recent studies have shown that SMRT DNA sequencing can be used to directly identify diverse modified nucleotides in synthetic templates and plasmids[22,23]. In SMRT sequencing, the time required for the incorporation of each nucleotide can be monitored, in addition to the specific base selected. Kinetic variation (KV), or variation in the rate at which DNA polymerase incorporates bases into DNA during synthesis, is highly correlated with the presence of modifications in the template[23,24]. Furthermore, different modifications, such as m6A and m5C, are linked to distinct kinetic profiles[22–24]. Therefore, SMRT sequencing offers the possibility of directly detecting chemical modification (or damage) to nucleotides and discriminating between such changes[22,23]. In addition, because KV is detected in a strand-specific fashion at the level of single molecules, the percentage of DNA strands with a particular modification at each site in the genome can be estimated[24].

Although the capacity for DNA methylation seems to be widespread throughout the bacterial and archeal kingdoms, the extent and functional consequences of DNA methylation in bacteria have not been extensively studied. Most bacterial methyltransferases (MTases) are components of RM systems. In such systems, MTases are partnered with an associated restriction endonuclease that can cleave a DNA at a target sequence, but only if it has not been methylated[3,25]. The primary role of RM systems is to defend host cells from invasion by foreign DNA. However, there is some evidence that the methylation catalyzed by RM MTases can affect other processes at the level of transcription

[1]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota, USA. [2]Division of Infectious Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [3]Howard Hughes Medical Institute, Boston, Massachusetts, USA. [4]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA. [5]Pacific Biosciences, Menlo Park, California, USA. [6]Department of Statistics, Stanford University, Stanford, California, USA. [7]Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, China. [8]Department of Automation, Tsinghua University, Beijing, China. [9]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, USA. [10]New England Biolabs, Inc., Ipswich, Massachusetts, USA. [11]These authors contributed equally to this work. Correspondence should be addressed to E.E.S. (eschadt@pacificbiosciences.com) or M.K.W. (mwaldor@rics.bwh.harvard.edu).

Received 19 July; accepted 22 October; published online 08 November 2012; corrected after print 3 May 2013; doi:10.1038/nbt.2432

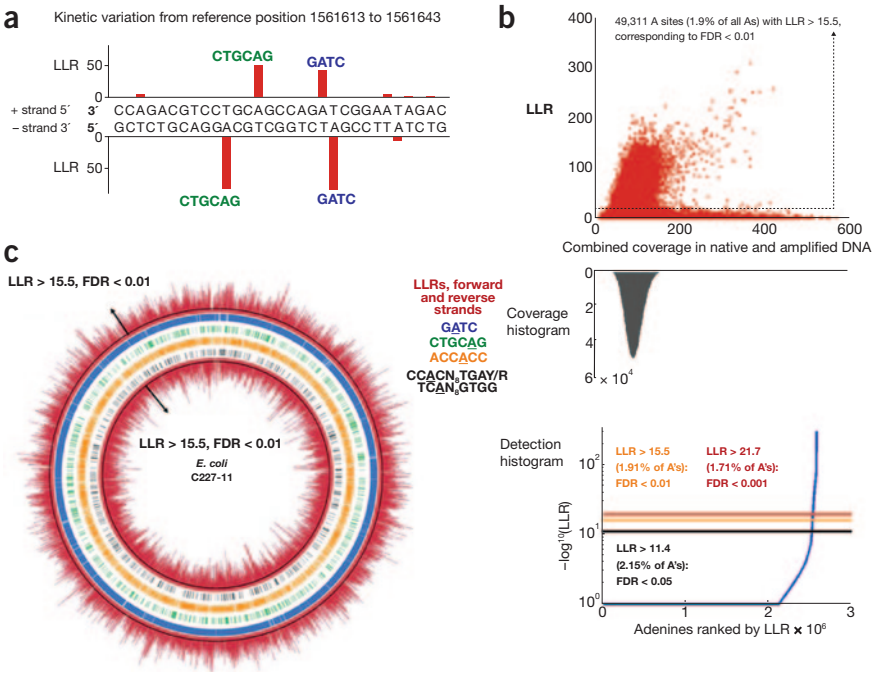

Figure 1 Extensive kinetic variation detected in C227-11. (**a**) Representative KV events detected at two loci in the C227-11 genome. For each position the loglikelihood ratio (LLR) statistic is plotted and represents the likelihood that the enzyme kinetics at the indicated position in the native C227-11 DNA is significantly ($P < 2.58 \times 10^{-8}$; FDR < 1%) different than the kinetics at the corresponding position in whole genome amplified C227-11 DNA. The higher the LLR, the more support exists for KV events. The only KV signals in this region are at the indicated positions and correspond to CTGCAG and GATC motifs on both the plus and minus strands. (**b**) Scatter plot and histogram of the extensive KV events detected at adenines detected across the C227-11 genome (see **Supplementary Table 2** for summaries of the detections at G, C and T bases). The *x* axis represents the amount of sequence coverage supporting the tested position. At a 1% FDR (corresponding to an LLR of 15.5; dotted horizontal line) 49,311 adenines were detected as KV events (1.9% of all adenines). (**c**) Genome-wide detections of KV events at adenines identified in **b** are represented around the C227-11 chromosome as a circos plot. The LLRs of adenine sites on the positive and negative strands are plotted as the red curve on the outer and inner loops of the Circos plot, respectively. We subsampled 50,000 sites on each strand for illustration purposes. The black circles represent the 1% FDR line (LLR > 15.5). Blue, green, orange and black hash marks correspond to the locations of GATC, CTGCAG, ACCACC and CCACN$_8$TGAY/RTCAN$_8$GTGG motifs, respectively.

regulation[26], as has been found for 'orphan' MTases, which lack an associated partner restriction enzyme. Most investigations into the phenotypes associated with DNA methylation in bacteria have focused on the study of these orphan MTases (e.g., Dam and CcrM)[14,27], but these analyses have not assessed genome-wide methylation patterns.

We took advantage of the KV data that can be obtained using SMRT DNA sequencing to comprehensively detect 49,311 m6A residues, a frequent genome modification in bacteria, at single-nucleotide resolution in the whole genome of hemolytic uremic syndrome (HUS)-linked *E. coli* O104:H4 (**Supplementary Table 1**)[28]. We also detected 1,407 putative m5C residues. Without a priori knowledge of MTase specificity, we used these data to deduce the sequence motifs recognized by the MTases present in this strain, and correlated methylation patterns with transcriptional profiles, gene amplification and bioinformatics analysis of genes that were within 500 bases of the MTase-targeted sequences. The findings of this study indicate that RM system MTases have additional roles beyond protecting host genomes from foreign DNA.

RESULTS

Genome-wide identification of base modifications

SMRT DNA sequencing permits identification of modified template nucleotides, such as m6A and 5mC, because the rate of DNA synthesis by DNA polymerase is extremely sensitive to these modifications. The time needed for incorporation of a nucleotide at a particular site (indicated by the interpulse duration) can vary depending on whether the template is native DNA (contains modifications) or if the DNA has been subjected to whole-genome amplification (WGA DNA; no modifications). Recent analyses revealed that adenine-linked and cytosine-linked KVs are strongly associated with the presence of m6A and m5C, respectively, in plasmid templates[22]; but KV has not been analyzed genome-wide.

We evaluated polymerase kinetics for 10,453,550 of 10,557,800 (99%) positions in the genome of the O104:H4 serotype *E. coli* outbreak strain C227-11. The remaining 104,250 positions were not evaluated because they did not have sufficient coverage (≥5× sequence coverage in the native and WGA data sets). For each site, we tested whether the rate distributions for nucleotide incorporation using native and WGA templates were sampled from a single kinetic rate distribution or from two distinct distributions (see **Fig. 1a** for adenine detections)[24]. The vast majority of sites were not detected as KV events at a 1% false discovery rate (FDR), suggesting that KV is absent and that these sites are likely unmodified, whereas a small subset of sites were detected as KV events with high confidence (**Supplementary Table 2** and see **Fig. 1b** for adenine detections). Permutation testing resulted in the detection of 1.9% sites (51,972 of 10,453,550 bases tested) at an FDR < 1%. Of the 51,972 bases detected, 49,311 (94.9%) were A, 1,407 (2.7%) were C, 833 (1.6%) were G and 421 were (0.8%) T. Whereas adenines were >4-fold enriched as KV sites, cytosines, guanines and thymines were 9.4, 15.8 and 30.5 times less likely to be detected as KV sites, respectively, than we would expect by chance (all Fisher's exact test, FET, *P* values for these under and over enrichments were $P < 10^{-300}$). These sites were distributed around the chromosome and exhibited marginal strand-specific biases (**Fig. 1c**).

We analyzed the local sequence contexts for all KV sites to assess whether modifications were randomly distributed among the bases in the genome or whether they were concentrated in particular sequence motifs. We used MEME[29] in a two-phase approach to search for motifs that were enriched for modifications. First, we searched for motifs of all lengths without gaps and identified four motifs, G<u>A</u>TC, CTGC<u>A</u>G, ACC<u>A</u>CC and C<u>C</u>WGG (detected methylated position is underlined), that were significantly enriched for KV events (FET, $P < 10^{-15}$; **Fig. 2** and **Table 1**). We found that 39,324 (94.1%) of the 41,791 adenines in analyzed GATC sites, 2,391 (96.2%) of the 2,486 adenines in analyzed CTGCAG sites and 4,130 (93.2%) of the 4,432 internal adenines in analyzed ACCACC sites had distinct kinetic profiles (FDR < 1%) in native DNA versus WGA DNA (Fisher's exact test, $P < 10^{-300}$ for all cases; **Table 1**). For the CCWGG motif, only 92 (0.34%) of the 26,938 occurrences were detected as modified (FDR < 1%), perhaps reflecting the low signal-to-noise ratio with the

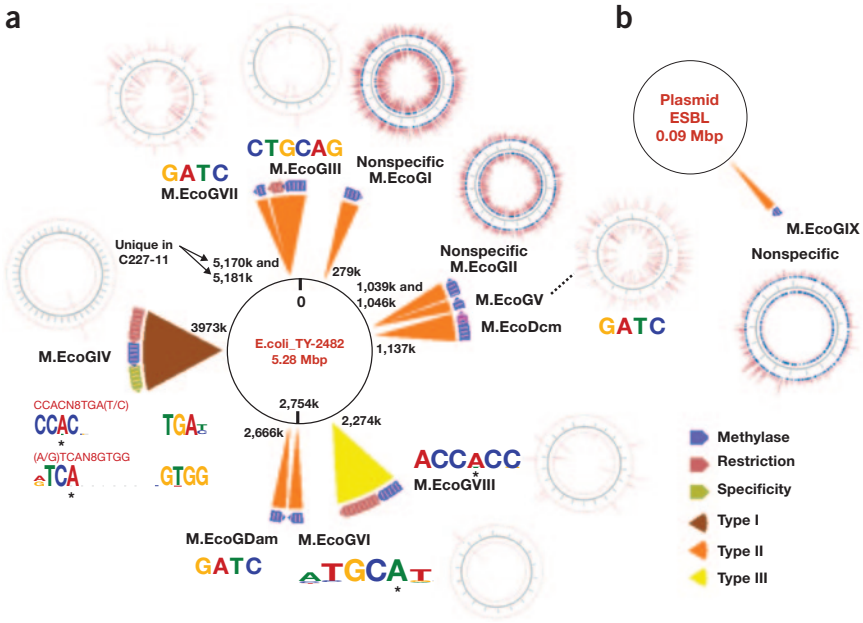

Figure 2 Identification and annotation of the MTases targeting the different sequence motifs in the C227-11 genome. (**a**,**b**) Locations of the nine predicted MTases (**Table 1**) are shown around the C227-11 chromosome (**a**) and ESBL plasmid (**b**). The triangle expansions at each MTase site are color-coded according to type, and the genes represented in each MTase region are also color-coded by gene type. Each MTase gene was cloned into plasmid pRRS and expressed in a methylase-free strain of *E. coli*. The plasmid constructs for each of the MTases are depicted as a circos plots at each of the MTase locations, with the inside of the annulus representing the coordinates of the plasmid, the blue hash marks representing the locations of adenines contained in the corresponding motif targeted by the MTase, and the two red curves around the plasmid representing the –log10(*P* value) for the LLR model for the two DNA strands. Associated with each circos plot is a Weblogo plot based on the top adenine sites detected in each case (FDR <0.1%). The adenines in all of the contexts indicated were detected as modified at an FDR <0.1%.

PacBio RS instrument when detecting m5C modifications[23]. Second, we ran MEME for all sites that displayed KV that was not explained either by the presence of motifs (GATC, CTGCAG, ACCACC and CCWGG) or by the secondary peaks that were associated with GATC modifications[23]. In this second stage of MEME analysis, gaps were allowed. This analysis yielded two complex, complementary sequence motifs, CCACN8TGA(T/C) and (A/G)TCAN8GTGG, that were substantially enriched for adenine modifications (in the form of CC^{m6}ACN8TGA(T/C) and (A/G)TC^{m6}AN8GTGG). We detected KV (FDR < 1%) at 440 of the 473 (93.0%) adenines in analyzed CCACN8TGA(T/C) sites and 449 (94.9%) of the 473 adenines in analyzed (A/G)TCAN8GTGG sites. Along with the three motifs detected in the previous analysis, these sites accounted for ~95% of all the adenine-linked KV detected. As GATC and CTGCAG, which contain 89% of all sites of adenine-linked KV detected, have previously been shown to be targeted by adenine MTases (Dam and PstI methylase, respectively), these data provide initial confirmation that the KV we observed reflects adenine methylation in the C227-11 genome.

Target sequences of the C227-11 MTases

Bioinformatic analysis of the C227-11 genome revealed that the strain encodes nine putative adenine-specific MTases, in addition to Dam (M.EcoGDam) and a single putative cytosine MTase (M.EcoGX) (**Fig. 2**, **Table 1** and **Supplementary Fig. 1**). For seven of the nine adenine-specific MTases, the target sites were unknown. One of the two remaining MTases, M.EcoGVI, corresponds to a previously characterized orphan MTase, YdhJ, that targets ATGCAT[30]. There are also two *dam* paralogs, one of which encodes M.EcoGV and has most similarity to M.EfaBMDam (43% identity), and the other encodes M.EcoGVII, which shows most similarity to M.EcoT1Dam (23% identity). Both paralogs seem to be encoded by uncharacterized prophages. We also identified a paralog of a PstI methyltransferase named M.EcoGIII, that had substantial similarity to M.PstI (42% identity) (**Table 1**). M.PstI is part of an RM system, whereas the *dam* paralogs lack associated partner endonucleases. It is possible that adenine-linked modifications of GATC sites in C227-11 reflect the activity of all three of these Dam-like enzymes, particularly because transcriptome analyses indicate that

Table 1 Summary of methyltransferases identified in the C227-11

Protein name	Number of gene copies	Chromosome coordinate	Observed specificity in plasmid	Active in outbreak strain	Number of occurrences in genome	Number of detections	Validated with restriction enzyme
M.EcoGI	2[a]	+279271	Nonspecific	No	NA	NA	NA
M.EcoGII	2[a]	–1040522	Nonspecific	No	NA	NA	NA
M.EcoGV	1	–1046318	G^{m6}AT/T^{m6}AC	G^{m6}ATC[b]	41,791	39,324 (94.1%)	G^{m6}ATC
M.EcoGVIII[c]	1	+2273609	ACC^{m6}ACC	ACC^{m6}ACC	4,432	4,130 (93.2%)	NA
M.EcoGVI	1	+2665879	ATGC^{m6}AT	No	2,000	0 (0%)	ATGC^{m6}AT
M.EcoGIV[c]	1	–3974814	CC^{m6}ACN8TGA(T/C)/ (A/G)TC^{m6}AN8GTGG	CC^{m6}ACN8TGA(T/C) and (A/G) TC^{m6}AN8GTGG	946	440 (93.0%)/ 449 (94.9%)	NA
M.EcoGIII[c]	1	–5184126	CTGC^{m6}A/ TGC^{m6}AG	CTGC^{m6}AG	2,486	2,391 (96.2%)	CTGC^{m6}AG
M.EcoGVII	1	–5170628	G^{m6}AT/T^{m6}AC	G^{m6}ATC[b]	41,791	39,324 (94.1%)	G^{m6}ATC
M.EcoGIX	1	+33659[d]	Nonspecific	No	NA	NA	NA
M.EcoGX	1	+1136734	Nonspecific	CCWGG	26,938	92 (0.34%)[e]	NA
M.EcoGDam	1	2754094	G^{m6}AT/T^{m6}AC	G^{m6}ATC[b]	41,791	39,324 (94.1%)	G^{m6}ATC

[a]M.EcoGI and M.EcoGII appear to be a duplication. [b]M.EcoGDam, M.EcoGV and M.EcoGVII are all methyltransferases that specifically target GATC; we cannot distinguish whether all or a subset target GATC in the outbreak strain. [c]These three methylases M.EcoGVIII, M.EcoGIV, M.EcoGIII have their corresponding restriction enzyme of type III, type I and type II, respectively. [d]The M.EcoGIX is in Plasmid TY1; the coordinate is given with respect to this plasmid. [e]Low detection rate owing to low signal-to-noise ratio of the m5C modifications on the PacBio RS sequencing instrument.
NA, not applicable.

Table 2 Comparison of adenine KV detection rates in CTGCAG-specific MTase–modified *E. coli* strains

Control strain (coverage)	Case strain (coverage)	DNA type	Number of A sites tested	Number of A sites with FDR <0.05	Number of CTGCAG sites tested	Number of CTGCAG sites with FDR <0.05	Fold enrichment of CTGCAG sites detected	Fisher's exact test *P* value[a]
C227-11 (28)	C227ΔRM (37)	Chromosome	2,558,210	2,646 (0.1%)	2,491	2,003 (80.4%)	777	$<10^{-300}$
C227-11 (25)	C227ΔRM (37)	Phage	14,569	8 (0.06%)	10	6 (60.0%)	1,093	4.4^{-19}
C227-11 (14)	C227ΔRM (10)	Plasmid 1	20,019	2 (0.01%)	64	1 (1.6%)	156	6.4^{-3}
C227-11 (11)	C227ΔRM (12)	Plasmid 2	18,464	13 (0.07%)	73	10 (13.7%)	195	1.4^{-22}
C227-11 (11)	C227ΔRM (12)	Plasmid 3	762	0 (0.0%)	0	N/A	N/A	N/A
K12 + ϕ104 (28)	K12 + ϕ104 RM deletion (40)	Chromosome	2,269,946	2361 (0.1%)	1,905	1,713 (90.0%)	865	$<10^{-300}$
K12 + ϕ104 (245)	K12 + ϕ104 RM deletion (26)	Phage	14,754	10 (0.07%)	8	7 (87.5%)	1,291	3.2^{-23}
K12 (48)	K12 + ϕ104 (40)	Chromosome	2,221,100	1,941 (0.09%)	1,859	1,807 (97.2%)	1,112	$<10^{-300}$

[a]The Fisher exact test assesses whether the number of CTGCAG sites detected was greater than expected by chance.

all three genes are expressed (**Supplementary Fig. 2**). Dam-dependent DNA methylation has previously been shown to modulate chromosome replication, DNA repair and transcription in *E. coli*[31]; however, genome-wide, strand-specific analysis of Dam methylation sites has not been previously carried out.

To confirm the activities and target sequence specificities of the nine uncharacterized, putative adenine MTases, we expressed a plasmid-borne gene encoding each enzyme in a previously described MTase-free strain of *E. coli,* then performed SMRT sequencing of the plasmid DNA to determine the modification pattern (**Fig. 2**, **Table 1** and **Supplementary Figs. 1**, **3**–**8**)[22]. Characterization of MTase targets in plasmid systems could be difficult to interpret, as overproduction of the MTases in such systems could result in off-target activity and underproduction might result in inefficient modification of the target sequence. However, our SMRT analysis combines motifs detected as enriched by the KV analysis and then links specific MTases to the enriched motifs via the plasmid data. Adenine-specific modification of GATC sites was detected when the *dam* homologs were expressed in the plasmid system, but such modifications were also observed in GAT and ATC sites, potentially reflecting off-target activity of the MTase from the plasmid expression system as has been noted previously (**Supplementary Figs. 3** and **4**)[22]. Similarly, the PstI-like MTase M.EcoGIII was linked to modification of CTGCAG, in addition to CTGCA and TGCAG (**Supplementary Fig. 5**). We expressed a plasmid-borne gene encoding the single cytosine MTase enzyme M.M.EcoGX in an MTase-free strain of *E. coli* and then carried out a restriction digestion analysis on the plasmid DNA using a methylation-sensitive enzyme, to confirm that cytosine-linked KV in the CCWGG context reflected the presence of m5C (**Table 1**). Similarly, we expressed plasmid-borne genes encoding M.EcoGI, M.EcoGII and M.EcoGIII MTases in the MTase-free strain of *E. coli* and subjected the plasmid DNA to a restriction digest, the results of which confirmed the presence of m6A and that GATC (for M.EcoGV and M.EcoGVII) and CTGCAG (for M.EcoGIII) are the recognition motifs for these MTases. Notably, we also identified an MTase (M.EcoGVIII) with a putative target sequence of ACCACC (**Fig. 2a**), which likely accounts for formation of $ACC^{m6}ACC$ in the C227-11 DNA. M.EcoGVIII seems to be part of a type III RM system, as it modifies just one strand of the recognition sequence and is flanked by a close homolog of type III restriction enzyme genes. Similarly, the plasmid-based analyses indicate that M.EcoGIV (part of a typical type I RM, with the usual genes coding for a restriction endonuclease, methyltransferase and specificity subunit) generates 5′-$CC^{m6}AN_8TGA$(T/C)-3′ and 5′-(A/G)$TC^{m6}AN_8GTGG$-3′, complementary motifs that we also identified in the genomic analyses of C227-11 DNA (**Table 1** and **Fig. 2a**).

Additionally, overexpression of M.EcoGI, M.EcoGII, M.EcoGVI and M.EcoGIX produced KV signatures that were not detected by SMRT sequencing of the C227-11 genome. As expected, the enzyme M.EcoGVI targeted the second adenine in ATGCAT sites (**Fig. 2a** and **Supplementary Fig. 6**), and the other three MTases, M.EcoGI, M.EcoGII and M.EcoGIX, had nonspecific activity (**Fig. 2a** and **Supplementary Figs. 7** and **8**). As RNA-seq analysis of C227-11 indicated that several of these enzymes are transcribed (**Supplementary Fig. 2**), it seems likely that the production and/or activity of these nonspecific enzymes are subject to distinct modes of control in C227-11 that remain to be determined.

Although roughly 95% of the observed sites of adenine modification in C227-11 were experimentally validated target sequences of the MTases present in the genome, we explored the possible sources of the remaining 2,577 sites of KV. Just over 44% (1,143) of these KV events are likely to be secondary peaks that resulted from a separate primary KV event, given that modification of one base may cause variation in the kinetics at nearby base locations[22–24]. Off-target effects likely explain many of the remaining 1,434 sites of KV[22]. For example, we detected 333 sites contained in motifs that had a single-base difference from the GATC motif, a 1.45-fold enrichment over what would be expected by chance (FET $P < 10^{-12}$), and 727 sites contained in motifs that had a two-base difference from GATC (1.17-fold enrichment; FET $P < 10^{-8}$). These enrichments might indicate that 74% (1,060 sites) of the remaining 1,434 sites are likely the result of off-target effects, although we cannot exclude other explanations such as complex KV signatures in some contexts. The remaining 374 sites represent only 0.76% of the total, and given that these detections were made at a 1% FDR, this number is well within the range of the expected number of false positive events. Therefore, we conclude that almost all of the adenine sites with KV are the result of modifications by m6A MTases.

Detection of partially methylated and unmethylated sites

Unlike restriction enzyme digests, SMRT sequencing can be used to identify loci at which only one of the two DNA strands is methylated. We analyzed the extent of methylation associated with several motifs including GATC, CTGCAG, ACCACC, CCWGG, $CC^{m6}ACN_8TGA$(T/C) and (A/G)$TC^{m6}AN_8GTGG$. Of the GATC sites that we detected as methylated, 1,966 (9.4%) were detected as methylated on only one strand, and for 220 GATC sites (1.1%) we detected neither strand as methylated (**Supplementary Table 2**). For the CTGCAG and CCACN8TGA(T/C) motifs, 75 (6%) and 34 (7.2%) of the sites were detected as hemimethylated and 9 (0.7%)

and 0 (0.0%) with neither strand detected as methylated, respectively (**Supplementary Table 2**). However, these findings are likely to overestimate the frequency of hemimethylation and nonmethylation. Declaring a locus as truly unmethylated requires more stringent criteria than the detection of modification events, given that lack of detection of a modification event may reflect a lack of power (that is, a lack of coverage) to make the detection, not a lack of methylation at the site. For most of the 2,406 GATC sites in the C227-11 genome that were denoted as unmodified, we could not confirm these findings because of a lack of sufficient sequence coverage. A minimum of 28× coverage is needed to ensure that the probability (power) that we detect a modification event (at the 0.1 significance threshold) at a site, if the site is truly modified, is > 99.99%. However, for 23 GATC sites we were >99.99% powered to make the KV detection, but did not, so we can predict these sites to be unmethylated with high confidence. Of the 23 unmethylated detections, 20 reflected reverse complement pairs in the same GATC locus, therefore representing fully unmethylated sites, and the remaining three correspond to confidently hemimethylated loci. All 23 sites were in noncoding regions, which is a significant bias (Fisher exact test $P = 1.3 \times 10^{-6}$) that is consistent with a previous report[11,32]. The genes nearest to, or containing, these sites were enriched for the phosphotransferase system ($P = 4 \times 10^{-5}$) pathway genes.

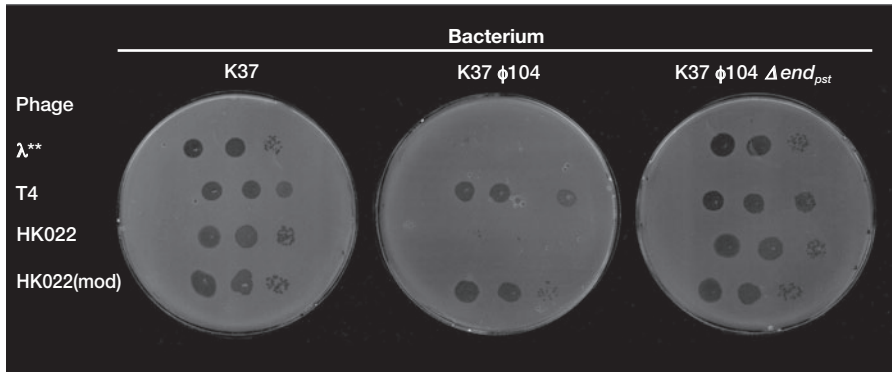

Figure 3 ϕ104 encodes a functional restriction-modification system. Ten microliters of serial 100-fold dilutions of the indicated phage were spotted (left to right) on top agar-seeded lawns of the *E. coli* strain K12/K37 and lysogenic derivatives thereof. λ** has the immunity of lambdoid phage 434 (ref. 50), but is primarily λ. The modified lysate, HK022(mod), was obtained by growing lambdoid phage HK022 (ref. 51) in the lysogen of K12/K37 that carries a ϕ104 prophage with a deletion substitution of the Pst1 endonuclease gene (K37ϕ104Δend_{pst}).

Even for loci identified as modified, it is likely that KV will not always be detected in all sequencing reads corresponding to the modified site because methylation is not simultaneous with DNA replication. However, we explored whether any modified sites had a higher proportion of reads lacking evidence of KV, as such heterogeneity might have important biological meaning. We analyzed KV data for a high-coverage subset (>75-fold coverage) of modified GATC sites using an unsupervised mixture model[24], to detect sites that yielded two distinct interpulse-duration distributions. Of the 3,495 GATC sites analyzed that were not within ten bases of other modified sites, 1,361 (38.9%) were predicted to be partially methylated (FDR of ~1%), whereas 2,584 (61.1%) were fully modified. Biological processes such as replication and tight binding proteins such as repressors can prevent methylation; other processes are also known to be regulated by hemimethylation[27,33–36].

To confirm the methylated and unmethylated sites identified using SMRT sequencing, we performed restriction enzyme and PCR-based analyses of methylation status for a subset of these sites. We randomly selected five of the top 20 GATC sites detected as fully unmodified with the highest power for detection but with nonsignificant *P* values, and two of the top 20 GATC sites detected as fully modified with the lowest *P* values. For each site tested, the PCR-based analyses yielded results consistent with those from the analyses of KV (**Supplementary Fig. 9**).

Host processes impacted by the phage-encoded RM system

The MTase targeting CTGCAG, M.EcoGIII, is part of an RM system encoded in the same prophage (ϕStx104) genome as *stxAB*, the genes encoding Shiga toxin. This phage is not present in most O104:H4 *E. coli* strains; its acquisition was a key event in the evolution of the O104:H4 outbreak strain, as Shiga toxin is the primary cause of HUS[37]. Notably, SMRT analyses of eight non-outbreak isolates of O104 *E. coli* revealed no sign of KV associated with CTGCAG sites (**Supplementary Figs. 1** and **10**). Furthermore, a derivative of C227-11 lacking M.EcoGIII, along with its associated restriction endonuclease R.EcoGIII, (referred to here as C227ΔRM) had a kinetic variation profile that differed from that of C227-11, specifically at CTGCAG sites. Among adenines whose modification status differed between C227-11 and C227ΔRM, there was a >700-fold enrichment for nucleotides in CTGCAG motifs ($P < 10^{-300}$; **Table 2**). The frequency with which other modified motifs (GATC, ACCACC, CCACN8TGA(T/C) and (A/G)TCAN8GTGG) were detected did not vary between these strains (data not shown). These data confirm the plasmid-based analyses that we carried out, and indicate that CTGCAG modification in C227-11 is most likely catalyzed by M.EcoGIII.

To explore the impact of the PstI-like RM system, RM.EcoGIII, we transduced a marked (*stxAB*<>*gntR*) version of ϕStx104 into *E. coli* K-12 strain K37. SMRT sequencing analysis of the wild-type *E. coli* K-12 strain K37, the *E. coli* K-12 strain K37ϕStx104 lysogen, and an isogenic derivative of the lysogen lacking the RM.EcoGIII-encoding genes revealed that the presence of the RM genes was linked to KV at adenines in CTGCAG sites, which were >800-fold enriched ($P < 10^{-300}$) relative to all adenines displaying different modification profiles for each comparison (**Table 2**). No other motifs showed modification (KV) that was linked to the presence of this RM system (data not shown). Additionally, we found that the DNA modifications present in the K37ϕStx104 lysogen render this lysogen resistant to infection by other lambda–like phages such as λimm434 and HK022 (**Fig. 3**) but not to T4, a phage that is often resistant to endonuclease digestion because its genome is highly modified[38]. The resistance of the *E. coli* K-12 strain K37ϕStx104 to infection by λ (as shown by the λ variant λimm434) and HK022 was dependent on the presence of the MTase-associated endonuclease (**Fig. 3**). However, we did not observe exclusion when we tested a derivative of HK022 (HK022(mod)), whose DNA was modified by growth in a strain expressing M.EcoGIII carried by ϕStx104. Together, these observations show that the PstI-like RM system, RM.EcoGIII, in ϕStx104, is functional after transduction to *E. coli* K12/K37.

The presence of the PstI-like RM system, RM.EcoGIII, also influences the growth of *E. coli*. The K37ϕStx104 lysogen had a reduced rate of growth compared with that of the parental strain, which was alleviated by deletion of the endonuclease (R.EcoGIII) and methyltransferase M.EcoGIII from ϕStx104 (**Supplementary Fig. 11**). In contrast, deletion of RM.EcoGIII in *E. coli* C227-11 had the opposite effect, resulting in a

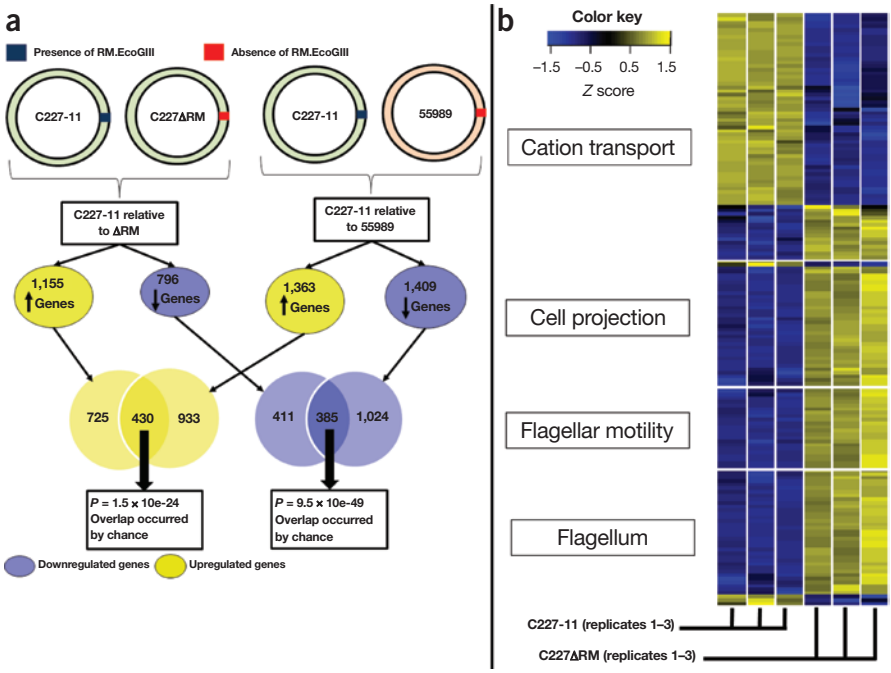

Figure 4 The RM system associated with M.EcoGIII regulates the expression of many genes and pathways. (**a**) Several genes are identified as differentially expressed at the 1% FDR level in C227-11 relative to C227ΔRM and in C227-11 relative to 55989. Significant overlap between the signatures is observed, likely reflecting common effects of the absence of this RM system. These signatures are also enriched for common pathways (**Supplementary Table 5**). (**b**) Heat map of 109 genes that are differentially expressed in three replicates of C227-11 relative to three replicates of C227ΔRM (**Supplementary Table 8**), and that are in the pathways indicated in **Supplementary Table 5** (cation transport, cell motility, cell projection and/or flagellum). Genes that are downregulated (upregulated) in C227-11 relative to C227ΔRM are colored different shades of blue (yellow) depending on the degree of downregulation (upregulation). The *Z* score reflects the degree of down (*Z* score < 0) or up (*Z* score > 0) regulations, computed by subtracting the mean of the log transformed expression values and dividing by the s.d. for each gene over all samples scored.

reduced growth rate (**Supplementary Fig. 12**). This disparity suggests that the *E. coli* O104:H4 host has adapted to the presence of ϕStx104.

We also observed amplification of fragments of the C227ΔRM genome and that might explain this strain's reduced growth rate. Comparisons of the depth of SMRT sequence data revealed that four genomic regions, spanning several thousand bases, had a maximum of greater than three-fold more coverage of DNA sequencing reads in C227ΔRM compared to the wild-type parental strain C227-11 (**Supplementary Fig. 13**). We tested these four regions for enrichments of pathways and gene classifications, and found that three of the four amplified peaks, 1, 2 and 7 (but not 6), were 12.5-fold (Fisher's exact test $P = 1 \times 10^{-2}$), 16.8-fold ($P = 6 \times 10^{-4}$) and 20.6-fold ($P = 2 \times 10^{-8}$) enriched, respectively, for phage-associated genes, suggesting that these regions correspond to fragments of prophages. Furthermore, these three regions were 5-fold (Fisher's exact test $P = 8 \times 10^{-6}$), 5.5-fold ($P = 2 \times 10^{-6}$) and 4.8-fold ($P = 2 \times 10^{-9}$) enriched, respectively, for CTGCAG sites modified in C227-11. In addition, we identified reads consistent with excision and circularization of the amplified peak 7 region (**Supplementary Fig. 14**); however, it is not clear whether the circular products are a cause or consequence of amplification. Whether these observations represent replication of these prophage-borne loci, controlled in some way by the RM system, remains to be shown, although the data are consistent with ϕStx104 being intimately co-adapted to at least some of the other mobile elements in the C227-11 genome.

The ϕStx104 RM system modulates transcription

To study the functional consequences of methylation of CTGCAG sites, we compared the transcriptomes of C227-11 and C227ΔRM. Of the 5,131 genes annotated in the C227-11 genome (**Supplementary Table 3**), 1,951 (~38%) were differentially expressed between these strains at a 1% FDR (**Supplementary Table 4**), with effect sizes ranging from a log base 10 ratio < −1.2 (downregulated) to >1.2 (>15-fold upregulated) in C227-11 compared to C227ΔRM. There were 1,155 genes with increased transcript abundance in C227-11 compared to C227ΔRM (including the RM.EcoGIII-encoding genes expressed in C227-11 but with no reads detected in C227ΔRM) and 796 genes with reduced transcript abundance (**Fig. 4a**). The set of genes with elevated expression was markedly enriched for several related functional Gene Ontology (GO) terms, including transition metal ion transport ($P = 1.6 \times 10^{-16}$), di-/tri-valent inorganic cation transport ($P < 3 \times 10^{-15}$), metal ion transport ($P = 5 \times 10^{-12}$) and cation transport ($P = 8.4 \times 10^{-10}$) (FDR < 5% in all cases). The genes with reduced expression were also significantly enriched for several functional GO terms, including flagellum ($P = 3.4 \times 10^{-14}$), cell projection part ($P < 4.6 \times 10^{-11}$), cell motility ($P = 7.4 \times 10^{-9}$) and flagellar motility ($P = 7.4 \times 10^{-9}$) (**Fig. 4b** and **Supplementary Table 5**). Notably, cell projection and flagellum gene sets were also enriched for genes within 500 bases of methylated CTGCAG sites ($P = 3.9 \times 10^{-7}$ and $P = 4.1 \times 10^{-5}$, respectively), suggesting that CTGCAG modifications may directly influence gene expression in these pathways. We also found that C227ΔRM had reduced motility in soft agar (**Supplementary Fig. 15**), suggesting that the alterations in gene expression are physiologically meaningful.

We also compared the transcriptomes of C227-11 and a closely related *E. coli* O104 strain, 55989, which does not contain the M.EcoGIII-bearing ϕStx104 (**Supplementary Table 6**). We observed significant overlaps between the genes upregulated in each comparison (1.5-fold enrichment, Fisher's exact test $P < 10 \times 10^{-30}$) as well as overlaps in those downregulated (1.9-fold enrichment, Fisher's exact test $P < 10 \times 10^{-30}$), suggesting that the CTGCAG modifications have similar consequences in these genetic backgrounds (**Fig. 4a**). As an additional confirmation of the pathways influenced by the RM.EcoGIII system, we compared the transcriptomes of the common laboratory strain *E. coli* K12/K37 and K37ϕStx104 (**Supplementary Table 7**). As with the other comparisons, the presence of ϕStx104 and in particular the gene M.EcoGIII contained in this phage was associated with increased transcript abundance for genes in the cation transport pathways and reduced transcript abundance for genes in the flagellum and cell-projection pathways (**Supplementary Table 5**).

DISCUSSION

Our findings suggest that SMRT DNA sequencing has enormous potential to transform our capacity to simply, rapidly and comprehensively catalog epigenetic modifications of genomes. We demonstrated that analyses of the KV data obtained during routine DNA sequencing with the SMRT platform can be used for genome-wide detection of m6A

bases. This single molecule–based approach yields strand-specific information regarding the presence of methylation at each nucleotide, and it enables quantitative analyses of the frequency of methylation at each site. Furthermore, we show that it is possible to deduce the target sites of MTases that catalyze m6A modifications solely from analyses of KV data. Finally, the power of combining SMRT sequencing–based characterization of the methylome with additional approaches is illustrated by our finding that an MTase component of an RM system has substantial effects on bacterial gene expression and DNA replication; its role is not limited to protecting the genome against foreign sequences.

The combination of genome-wide analyses of m6A, gene expression studies and plasmid-based assays of individual MTases revealed several unexpected attributes of C227-11 MTases and suggests that SMRT sequencing could greatly enhance our understanding of the activities and roles of DNA MTases. We identified two new MTases in *E. coli* C227-11: M.EcoGVIII, a type III enzyme that generates ACC^{m6}ACC, and M.EcoGIV, a type I enzyme that generates 5′-CC^{m6}ACN_8TGA(T/C)-3′ and 5′- A/G)TC^{m6}AN_8GTGG-3′. Nearly all of the target sites for these two enzymes were methylated in the *E. coli* C227-11 genome. In addition, we found that *E. coli* C227-11 encodes several MTases (M.EcoGVI, M.EcoGI, M.EcoGII and M.EcoGIX) that are active when exogenously expressed but have low or no activity in the host strain. As C227-11 seems to transcribe the genes encoding these four proteins, additional studies to determine why their products' target sites are not routinely modified in this strain are warranted. M.EcoGI, M.EcoGII and M.EcoGIX all seem to lack sequence specificity but can instead modify adenines in a wide variety of sequence contexts. These three proteins are unrelated to a recently described family of adenine-specific MTases that can modify all adenines except for those that are present in homopolymer runs[39]. The observation that in a plasmid system M.EcoGIX seems to generate m6A in strand-specific fashion (**Supplementary Fig. 8**), warrants additional investigation as this activity has not been observed previously.

MTases that are components of RM systems are often associated with mobile genetic elements[40], and our work demonstrated that the effects of such RM systems can extend far beyond the mobile elements that harbor them. Lysogenization of a precursor of the *E. coli* O104:H4 strain that caused the large outbreak of HUS in Germany in the spring of 2011 with the lambdoid phage ϕStx104, which encodes both Shiga toxin and M.EcoGIII, not only rendered the strain capable of producing Shiga toxin, the cause of HUS but also markedly altered its transcriptome. The expression of more than one-third of C227-11 genes were significantly altered ($P < 0.006$) when the PstI-like RM system, RM.EcoGIII, was deleted from ϕStx104 (the C227ΔRM strain), although there was not a compelling correlation detected between m6A modification events in CTGCAG and differentially expressed genes, despite a number of pathways enriched for CTGCAG modifications being detected as differentially expressed (**Fig. 4b** and **Supplementary Table 5**). The widespread effect on the C227-11 transcriptome that was associated with deletion of the RM.EcoGIII system in ϕStx104, coupled with the modest growth defect of the C227ΔRM, suggests that the lysogenic conversion of *E. coli* O104:H4 strain was not a recent event. Furthermore, M.EcoGIII is associated with the apparent amplifications of fragments of three additional lambdoid prophages in C227-11 through mechanisms that remain to be defined.

The approach taken here to elucidate m6A and m5C marks in *E. coli* is just a first step in leveraging the extensive kinetic information that can be detected using SMRT sequencing. The ability of the third-generation SMRT sequencing technology to detect m6A-type modifications in DNA goes beyond the current capabilities of current-generation high-throughput sequencing technologies, opening the door to new discoveries such as those detailed in this report and to resolving long-standing questions regarding the functional consequences of DNA methylation. Additional improvements to SMRT sequencing for the analysis of bacterial genomes are likely to include refining the detection capability for different nucleotide types, enhancing the statistical modeling to extract more information out of the kinetic SMRT sequencing (to increase the power to make detections) and expanding capabilities to experimentally validate epigenetic changes. With additional advances we should be able to comprehensively characterize the potentially vast array of modification events that occur over whole bacterial genomes or genomes of higher organisms, thus enabling a fuller assessment of the functional role of DNA in defining the complexity of living systems.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The C227-11 assembly, GenBank: AFST00000000.1. The raw read data for C227-11 and K12/K37, SRA: SRA038239 and at the website http://www.pacbiodevnet.com/Share/Datasets/E-coli-Outbreak-Modifications. The assembly for TY-2482 is available at http://gigadb.org/e-coli/. The sequence data, target sequences and associated details on all methylase enzymes identified in this report have been deposited in the REBASE database (http://rebase.neb.com/rebase/rebase.html).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This study was supported in part by a US National Science Foundation grant IIS0916439 (G.F. and V.K.) and NIH R37 AI-42347 and HHMI (M.K.W.).

AUTHOR CONTRIBUTIONS

G.F., D.M., M.K.W. and E.E.S. designed the experiments; G.F., D.M., D.I.F., A.M., M.C.C., O.B., Z.F., I.A.M., A.K.-P., A.C., R.J.R., J.K., S.W.T., V.K., M.K.W. and E.E.S. designed the methods; D.M., D.I.F., A.M., M.C.C., M.C.M., O.J.J., G.D., T.A.C., K.L., I.A.M., A.K.-P. and A.C. carried out all sample-preparation experiments, all sequencing runs and all validation experiments; G.F., D.M., D.I.F., A.M., M.C.C., O.B., Z.F., B.L., I.A.M., B.M.D., A.K.-P., A.C., R.J.R., V.K., M.K.W. and E.E.S. jointly analyzed the data sets; and G.F., D.M., D.I.F., A.M., M.C.C., M.C.M., O.J.J., T.A.C., B.M.D., A.K.-P., A.C., R.J.R., J.K., M.K.W. and E.E.S. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Published online at http://www.nature.com/doifinder/10.1038/nbt.2432.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929–930 (2009).
2. Kumar, S. *et al.* The DNA (cytosine-5) methyltransferases. *Nucleic Acids Res.* **22**, 1–10 (1994).
3. Roberts, R.J., Vincze, T., Posfai, J. & Macelis, D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **38**, D234–D236 (2010).
4. Swinton, D. *et al.* Purification and characterization of the unusual deoxynucleoside, α-*N*-(9-β-D-2′-deoxyribofuranosylpurin-6-yl)glycinamide, specified by the phage Mu modification function. *Proc. Natl. Acad. Sci. USA* **80**, 7400–7404 (1983).
5. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
6. Warren, R.A. Modified bases in bacteriophage DNAs. *Annu. Rev. Microbiol.* **34**, 137–158 (1980).
7. Wyatt, G.R. & Cohen, S.S. A new pyrimidine base from bacteriophage nucleic acids. *Nature* **170**, 1072–1073 (1952).
8. Anway, M.D., Cupp, A.S., Uzumcu, M. & Skinner, M.K. Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science* **308**, 1466–1469 (2005).
9. Jirtle, R.L. & Skinner, M.K. Environmental epigenomics and disease susceptibility. *Nat. Rev. Genet.* **8**, 253–262 (2007).
10. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).
11. Casadesús, J. & Low, D. Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.* **70**, 830–856 (2006).

12. Collier, J., McAdams, H.H. & Shapiro, L. A DNA methylation ratchet governs progression through a bacterial cell cycle. *Proc. Natl. Acad. Sci. USA* **104**, 17111–17116 (2007).
13. 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).
14. 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).
15. Stephens, C., Reisenauer, A., Wright, R. & Shapiro, L. A cell cycle-regulated bacterial DNA methyltransferase is essential for viability. *Proc. Natl. Acad. Sci. USA* **93**, 1210–1214 (1996).
16. van der Woude, M., Braaten, B. & Low, D. Epigenetic phase variation of the pap operon in *Escherichia coli*. *Trends Microbiol.* **4**, 5–9 (1996).
17. Cokus, S.J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
18. Kahramanoglou, C. *et al.* Genomics of DNA cytosine methylation in Escherichia coli reveals its role in stationary phase transcription. *Nat. Commun.* **3**, 886 (2012).
19. Booth, M.J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).
20. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).
21. Schadt, E.E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum. Mol. Genet.* **19**, R227–R240 (2010).
22. Clark, T.A. *et al.* Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res.* **40**, e29 (2012).
23. Flusberg, B.A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465 (2010).
24. Schadt, E.E. *et al.* Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome Res.* advance online publication, doi:10.1101/gr.136739.111 (23 October 2012).
25. Roberts, R.J. & Halford, S.E. *Type II Restriction Enzymes* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 1993).
26. Srikhanta, Y.N. *et al.* Phasevarions mediate random switching of gene expression in pathogenic *Neisseria*. *PLoS Pathog.* **5**, e1000400 (2009).
27. Reisenauer, A., Kahng, L.S., McCollum, S. & Shapiro, L. Bacterial DNA methylation: a cell cycle regulator? *J. Bacteriol.* **181**, 5135–5139 (1999).
28. Rasko, D.A. *et al.* Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N. Engl. J. Med.* **365**, 709–717 (2011).
29. Bailey, T.L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202-8 (2009).
30. Broadbent, S.E., Balbontin, R., Casadesus, J., Marinus, M.G. & van der Woude, M. YhdJ, a nonessential CcrM-like DNA methyltransferase of *Escherichia coli* and Salmonella enterica. *J. Bacteriol.* **189**, 4325–4327 (2007).
31. Wion, D. & Casadesus, J. *N*6-methyl-adenine: an epigenetic signal for DNA-protein interactions. *Nat. Rev. Microbiol.* **4**, 183–192 (2006).
32. Low, D.A., Weyand, N.J. & Mahan, M.J. Roles of DNA adenine and methylation in regulating bacterial gene expression and virulence. *Infect. Immun.* **69**, 7197–1204 (2001).
33. Camacho, E.M. & Casadesus, J. Regulation of traJ transcription in the Salmonella virulence plasmid by strand-specific DNA adenine hemimethylation. *Mol. Microbiol.* **57**, 1700–1718 (2005).
34. Løbner-Olesen, A., Skovgaard, O. & Marinus, M.G. Dam methylation: coordinating cellular processes. *Curr. Opin. Microbiol.* **8**, 154–160 (2005).
35. Messer, W. & Noyer-Weidner, M. Timing and targeting: the biological functions of Dam methylation in *E. coli*. *Cell* **54**, 735–737 (1988).
36. Roberts, D., Hoopes, B.C., McClure, W.R. & Kleckner, N. IS10 transposition is regulated by DNA adenine methylation. *Cell* **43**, 117–130 (1985).
37. Kaper, J.B., Nataro, J.P. & Mobley, H.L. Pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.* **2**, 123–140 (2004).
38. Karam, J.D. & Drake, J.W. Molecular biology of bacteriophage T4 (American Society for Microbiology, Washington, DC, USA, 1994).
39. Drozdz, M., Piekarowicz, A., Bujnicki, J.M. & Radlinska, M. Novel non-specific DNA adenine methyltransferases. *Nucleic Acids Res.* **40**, 2119–2130 (2012).
40. Furuta, Y., Abe, K. & Kobayashi, I. Genome comparison and context analysis reveals putative mobile forms of restriction-modification systems and related rearrangements. *Nucleic Acids Res.* **38**, 2428–2443 (2010).

ONLINE METHODS

Predicting DNA MTases. Software modules combined with internal databases constitute the SEQWARE resource, a suite of software tools for managing, processing and annotating sequence data (http://seqware.github.com/). New sequence data are scanned locally for homologs that are already identified and for annotated RM systems in REBASE[3]. Sequence similarity from BLAST searches, the presence of predictive functional motifs and genomic context are the basic indicators of potential new RM system components. Heuristic rules, derived from knowledge about the gene structure of RM systems, are also applied to refine the hits. Attempts are made to avoid false hits caused by strong sequence similarity of RNA and protein methylases or hits based solely on nonspecific domains of RM enzymes, such as helicase or chromatin remodeling domains. SEQWARE then localizes motifs and domains, assigns probable recognition specificities, classifies accepted hits and marks Pfam relationships. All candidates are then inspected manually before being assigned as part of an RM system.

DNA samples and sequencing method. The isolation, DNA preparation and SMRT sequencing of the nine isolates described in **Supplementary Table 1** have been previously described[28]. To generate the kinetic control data set, WGA sequencing of the outbreak strain was performed. The sample was isolated and DNA extracted as previously described[28] from log-phase cells. An aliquot of ~25 ng of DNA was subjected to WGA using the REPLI-g Midi Kit (Qiagen) to erase DNA modifications. The WGA DNA was then sheared to an average size of ~300 base pairs via adaptive focused acoustics (Covaris).

With the WGA DNA prepared, sequencing libraries were prepared as previously described[41]. Briefly, sheared DNA was end-repaired and ligated to hairpin adaptors. Incompletely formed SMRTbells were degraded with a combination of exonuclease III (New England Biolabs) and exonuclease VII (USB). Primer was annealed and samples were sequenced on the PacBio RS as previously described[42,43].

RNA sequencing and analysis. RNA samples for three replicates each of the strains C227-11, C227ΔRM, K12/K37, K12-ϕStx104 and 55989 were treated with RiboZero bacterial Gram Negative ribosomal removal kit (Epicentre; MRZGN126). rRNA depleted RNA was chemically fragmented and then reverse transcribed with SuperScript III (Invitrogen) using random hexamers. The cDNA was converted to dsDNA, end-repaired, 5′ adenylated and then ligated to DNA adaptors enabling Illumina sequencing using the NEB Next mRNA-seq kit (New England BioLabs; E6100S). PCR for 16 cycles using KAPA HiFi HotStart DNA polymerase was performed to incorporate Illumina-specific sequences and molecular barcodes. RNA-seq libraries were sequenced on the Illumina HiSeq 2000 for a read length of 100 nt using standard methods[44].

Displaying and calling differentially expressed genes. We mapped raw RNA reads for each of the three replicates of strains C227-11, C227ΔRM and 55989 to the set of 5,380 genes predicted in the TY-2482 reference genome and for each of the three replicates of strains K12/K37 and K12-ϕStx104 to the set of 4,146 genes predicted in the K12/K37 reference genome. A gene was included for differential expression analysis if it had more than one count per million reads (CPM ≥ 1) in at least two samples. For the differential expression results depicted in **Figure 4** and listed in **Supplementary Tables 4**, **6** and **7**, the software program edgeR[45] was used to detect significantly differentially expressed genes ($P < 0.006$; FDR < 1%) for all groupings: C227-11 versus C227-ΔRM, C227-11 versus 55989, and K12/K37 versus K12-ϕStx104. To display the C227-11 and C227ΔRM expression data in **Figure 4b** (**Supplementary Table 8**), expression values were converted into a *Z* score by subtracting the mean of the log transformed expression values and dividing by the standard deviation for each gene over all samples scored.

Data processing. *SMRT sequence data.* Reads were processed and mapped using the BLASR mapper (http://www.pacbiodevnet.com/SMRT-Analysis/Algorithms/BLASR) and the Pacific Biosciences SMRTAnalysis pipeline (http://www.smrtcommunity.com/SMRT-Analysis/Software/SMRT-Analysis) using the standard mapping protocol. Interpulse durations (IPDs) were measured as previously described[23] for all pulses aligned to each position in the reference sequence. Reads from C227-11 were mapped to the TY-2482 genome reference sequence[28], and reads from K12/K37 were mapped to the K12 MG1665 genome reference sequence. All the raw SMRT sequencing data presented in this paper can be retrieved from http://www.pacbiodevnet.com/Share/Datasets/E-coli-Outbreak-Modifications.

Interpulse duration data filtering and interpulse duration normalization. We first removed IPDs that were above the 0.999 quantile to exclude extremely large values that are probably due to polymerase pauses rather than normal translocations. We then shifted all of the IPD values by adding 0.01 and then log-transformed the IPD values. This enabled us to make use of IPDs in log space that reflected fast polymerase translocations but that were not captured by the lowest resolution of the PacBio RS (0.01333 s, 75 frames s^{-1}). In log scale, IPD values were normalized across subreads by dividing the IPDS from each subread by the mean IPD of the corresponding subread. A minimum subread length of 10 was used to exclude subreads that had too few IPDs for the mean estimation procedure.

Base-level LLR test. A previously described single site likelihood model was used to detect kinetic variation events[24]. The likelihood under the null is nested in the likelihood under the alternative hypothesis. The null model assumes that the IPDs are sampled from a single rate distribution, whereas the alternative model assumes they are sampled from two different rate distributions. The likelihood model assumes that the filtered IPDs (log scale) are normally distributed. The test statistic is formed by finding the maximum likelihood estimates for the likelihoods under the null and alternative hypotheses, and then taking the LLR of the maximum likelihoods, which has a theoretical distribution that is chi-square with a single degree of freedom under the null hypothesis.

Permutation tests and the estimation of the FDR. We use permutation testing to estimate an empirical null distribution to calculate the FDRs. The permutation was done in such a way that, for each site, the permutation step shuffles the labels of an IPD being 'native' or 'WGA' while maintaining the number of total IPDs from the native and WGA data. For each permutation, we assigned a LLR for each of the 2,579,260 sites. With the LLRs for each site in the original (not shuffled) data and the LLRs for each site from P permutations, we can estimate the FDR of a site (with a LLR of l_i) by calculating the ratio between the fraction of LLRs of all the sites in all the permutations that are above l_i and the fraction of LLRs in the original data. We performed 20 permutations and found that the estimation of FDR was already stable.

Detecting confidently unmethylated GATC sites. To make confident calls of unmethylated GATC sites, we need to ensure a site not detected as a KV event is not simply explained as a false-negative event. For this purpose, we designed a power analysis on a plasmid dataset with 46 known methylated GATC sites. Specifically, we sampled each GATC site at different coverage and estimated the corresponding false negative rate based on the sampling of all the 46 GATC sites. We found that, for a GATC site, if both the native and WGA data have more than 27× IPDs available,

doi:10.1038/nbt.2432

only one in 50,000 GATC sites would have an LLR *P* value greater than 0.1, which corresponds to a false negative rate of 2×10^{-5}. Based on this, we confidently called a GATC site unmethylated if it has more than 27× IPDs in both the native and the WGA data yet has a LLR *P* value greater than 0.1. Given that there are 41,578 GATC sites in the filtered C227-11 dataset, the number of undetected GATC sites owing to expected false negative rate was only 0.83, that is, less than one.

Estimation of the FDR of partially methylated GATC sites. For this analysis we used the 3,495 apparently methylated GATC sites (FDR ≤ 0.01) that also had high coverage (more than 75×) in the native DNA sample and that did not have any neighboring bases detected as modified (10 bases upstream and downstream). To estimate the fraction of molecules modified at a given site, we applied an unsupervised mixture model[24] on the native IPDs, with the requirement that one of the IPD components (W) follows the same distribution as the IPDs in the WGA data at the same site. The statistical test then consisted of both estimating the fraction of molecules in component W and assessing whether the difference of the mean IPD in component W and the mean IPD in the other component (N) was significant, using the same LLR test used for detecting kinetic variation events. We identified 1,361 sites in which the 95% confidence interval for the estimated fraction of molecules in component N did not contain 1.0 and for which the mean IPD difference between components W and N was significant ($P < 0.05$); we declared such sites to be partially methylated. To estimate the FDR for these 1,361 partially methylated sites, we simulated native IPDs using the mean and standard deviation estimates from the original native IPDs, assuming there was no partial modification, and repeated this simulation on the 3,495 sites 100 times. Of the 349,500 simulation tests performed over the 100 random runs, only 1,949 sites met the above criteria for a partially methylated site, resulting in an FDR estimate of 0.014.

Motif-enrichment analysis. After each site in a genome is assigned with a LLR, we grouped these LLRs by motifs. Then, for each motif we used the nonparametric Wilcoxon's rank-sum test to compare the LLRs grouped into this motif with the overall LLRs (of the nucleotide of interest in the motif) as an indicator of the enrichment of this motif being methylated more than expected by chance. An indicator was associated with each test to reflect whether the LLRs associated with the motif were higher (1) or lower (0) than the background LLRs. We preferred this test over those requiring discretization of the base-level LLRs into two levels, given the LLRs for each base may not be reliable when the number of IPD observations upon which an LLR is based is insufficient. We found that directly testing the distribution of the LLRs associated with a motif with the background distribution had better power and robustness.

GO term enrichment analysis. For the GO term enrichment analysis, we used DAVID[46]. The background GO terms considered were the union of biological processes, cellular components and molecular functions, with the top five levels in the GO hierarchy excluded. An FDR cutoff of 0.01 was used to select enriched terms.

Isolation and cloning of genes encoding putative methyltransferases of *E. coli* C227-11. Individual MTase genes were amplified from *E. coli* C227-11 genomic DNA using Phusion DNA polymerase and inserted into the multicopy pRRS plasmid vector (via unique BamHI and PstI restriction sites) as previously described[22]. In the case of the putative type I restriction-modification system M.EcoGIV the genes encoding both the predicted and MTase and S (specificity) subunits were amplified and inserted into the plasmid vector as the latter confers sequence specificity for both endonuclease and methyltransferase activities in type I systems. Synthetic oligonucleotides used for amplification of M.EcoGI-GIII, (M+S)EcoGIV, M.EcoGV-GIX and M.EcoGDam are listed in **Supplementary Table 9**. Plasmid samples containing cloned MTase genes were propagated in strain ER2796, which lacks all *E. coli* MTase genes, before SMRT sequencing[22].

Constructing the C227ΔRM mutant. Deletion of the adjacent genes encoding the PstI-like methylase, M.EcoGIII and the corresponding endonuclease, R.EcoGIII (locus tags VBIEscCol192592_5132 and VBIEscCol192592_5131, respectively, in TY-2482) from the C227-11 genome was accomplished with standard allele-replacement techniques. A derivative of the *sacB*-containing vector pDM4 (ref. 47) harboring DNA regions flanking the *M* and *R* genes (pDM4-*MethEnd*) was introduced into C227 expressing M.EcoGIII from a pBAD18Km[48] derivative (pBAD18Km-Meth). After the deletion was confirmed by PCR, pBAD18Km-Met was eliminated from the mutant strain by serial passages in M9 minimal medium.

pDM4-MethEnd plasmid. The pDM4-MethEnd plasmid was constructed by amplification of DNA flanking regions 5′ *Meth* and 3′ *End* in *E. coli* O104:H4 strain C227-11. using primers pairs Fw-XbaI-Flk-Meth5′ and Rv-Flk-Meth5′-BglII-b, and Fw-BglII-Flk-Enz3′c and Rv-Flk-Enz3′-SacI, respectively. PCR products were annealed, extended by PCR amplification, digested with XbaI and SacI restriction enzymes and ligated in pDM4 vector digested with same enzymes.

pBAD18Km-Meth plasmid. pBAD18Km-Meth plasmid was constructed by DNA amplification of the M.EcoGIII gene in *E. coli* O104:H4 strain C227-11 with primers pairs Fw-XbaI-Sh-Meth and Rv-Meth-Xba. PCR product was digested with XbaI restriction enzyme and ligated into pBAD18Km digested with same enzyme.

Constructing the K37 lysogen with ϕStx104 prophage. A derivative of strain C227-11 in which the *stx2* genes were replaced with a gentamycin resistant cassette was treated with mitomycin C (2 μg ml^{-1}) to induce prophage production. The resulting lysate was clarified by centrifugation and the supernatant was spotted on an LB plate seeded with a lawn of K12 strain K37. After overnight incubation at 37°, bacteria from the turbid zone of lysis were struck on an LB-gentamycin plate and incubated overnight at 37°. Colonies from that plate produced phage that on infection of K37 produced progeny that were gentamycin-resistant. This observation along with PCR identification of M.EcoGIII confirmed that the lysogens harbored ϕStx104 prophage.

Gene replacements. To replace genes in the ϕStx104 prophage of C227-11 with antibiotic resistance cassettes we used a previously published method[49].

Primers. Fw-XbaI-Flk-Meth5′ (5′-gatccTCTAGAtgtaaaggtcggagatttcag); Rv-Flk-Meth5′-BglII-b (5′-gtcctgttcagatctaatcaaaAGTCTCATCATTCC); Fw-BglII-Flk-Enz3′c (5′-tttgattagatctgaacaggacCAGATAAATAGC); Rv-Flk-Enz3′-SacI (5′-cgggaGAGCTCGCCCCGACATCACACGCAGTC); Fw-XbaI-Sh-Meth (5′-gatcctctagaTAAGGAGGatttataaATGcttcaaaagctagac gtcgcag); and Rv-Meth-Xba (5′-tcgactctagaTCATgcagtcagagctcctagc).

41. Travers, K.J., Chin, C.S., Rank, D.R., Eid, J.S. & Turner, S.W. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* **38**, e159 (2010).
42. Chin, C.S. *et al.* The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* **364**, 33–42 (2011).
43. Korlach, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol.* **472**, 431–455 (2010).

44. Bentley, D.R. *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
45. Robinson, M.D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
46. Huang, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
47. Milton, D.L., O'Toole, R., Horstedt, P. & Wolf-Watz, H. Flagellin A is essential for the virulence of *Vibrio anguillarum*. *J. Bacteriol.* **178**, 1310–1319 (1996).
48. Guzman, L.M., Belin, D., Carson, M.J. & Beckwith, J. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J. Bacteriol.* **177**, 4121–4130 (1995).
49. Datsenko, K.A. & Wanner, B.L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. USA* **97**, 6640–6645 (2000).
50. Kaiser, A.D. & Jacob, F. Recombination between related temperate bacteriophages and the genetic control of immunity and prophage localization. *Virology* **4**, 509–521 (1957).
51. Dhillon, T.S. & Dhillon, E.K. Temperate coliphage HK022. Clear plaque mutants and preliminary vegetative map. *Jpn. J. Microbiol.* **20**, 385–396 (1976).

doi:10.1038/nbt.2432

Corrigendum: Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing

Gang Fang, Diana Munera, David I Friedman, Anjali Mandlik, Michael C Chao, Onureena Banerjee, Zhixing Feng, Bojan Losic, Milind C Mahajan, Omar J Jabado, Gintaras Deikus, Tyson A Clark, Khai Luong, Iain A Murray, Brigid M Davis, Alona Keren-Paz, Andrew Chess, Richard J Roberts, Jonas Korlach, Steve W Turner, Vipin Kumar, Matthew K Waldor & Eric E Schadt

***Nat. Biotechnol.* 30, 1232–1239 (2012); published online 8 November 2012; corrected after print 3 May 2013**

In the version of this article initially published, on p. 1235, line 32, the wrong MTases were given for the motif GATC. Instead of "…GATC (for M.EcoGI and M.EcoGII)…," it should have read, "…GATC (for M.EcoGV and M.EcoGVII)…." The error has been corrected for the PDF and HTML versions of this article.