730–743 *Nucleic Acids Research, 2016, Vol. 44, No. 2* *Published online 23 December 2015*
doi: 10.1093/nar/gkv1498

Precision methylome characterization of *Mycobacterium tuberculosis* complex (MTBC) using PacBio single-molecule real-time (SMRT) technology

Lingxiang Zhu[1,2,†], Jun Zhong[1,†], Xinmiao Jia[1,3,†], Guan Liu[4,†], Yu Kang[1,†], Mengxing Dong[1,3], Xiuli Zhang[1,3], Qian Li[1,3], Liya Yue[1,3], Cuidan Li[1,3], Jing Fu[1,3], Jingfa Xiao[1], Jiangwei Yan[1], Bing Zhang[5], Meng Lei[5], Suting Chen[4], Lingna Lv[4], Baoli Zhu[6], Hairong Huang[4,*] and Fei Chen[1,7,*]

[1]CAS Key Laboratory of Genome Sciences & Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [2]National Research Institute for Family Planning, Beijing 100081, China, [3]University of Chinese Academy of Sciences, Beijing 100049, China, [4]National Clinical Laboratory on Tuberculosis, Beijing Key Laboratory on Drug-resistant Tuberculosis Research, Beijing Chest Hospital, Capital Medical University, Beijing Tuberculosis and Thoracic Tumor Institute, Beijing 101149, China, [5]Core Genomic Facility, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [6]CAS Key Laboratory of Pathogenic Microbiology & Immunology, Institute Of Microbiology, Chinese Academy of Sciences, Beijing 100101, China and [7]Collaborative Innovation Center for Genetics and Development, China

Received July 31, 2015; Revised December 10, 2015; Accepted December 11, 2015

ABSTRACT

Tuberculosis (TB) remains one of the most common infectious diseases caused by *Mycobacterium tuberculosis* complex (MTBC). To panoramically analyze MTBC's genomic methylation, we completed the genomes of 12 MTBC strains (*Mycobacterium bovis*; *M. bovis* BCG; *M. microti*; *M. africanum*; *M. tuberculosis* H37Rv; H37Ra; and 6 *M. tuberculosis* clinical isolates) belonging to different lineages and characterized their methylomes using single-molecule real-time (SMRT) technology. We identified three ^{m6}A sequence motifs and their corresponding methyltransferase (MTase) genes, including the reported *mamA*, *hsdM* and a newly discovered *mamB*. We also experimentally verified the methylated motifs and functions of HsdM and MamB. Our analysis indicated the MTase activities varied between 12 strains due to mutations/deletions. Furthermore, through measuring 'the methylated-motif-site ratio' and 'the methylated-read ratio', we explored the methylation status of each modified site and sequence-read to obtain the 'precision methylome' of the MTBC strains, which enabled intricate analysis of MTase activity at whole-genome scale. Most unmodified sites overlapped with transcription-factor binding-regions, which might protect these sites from methylation. Overall, our findings show enormous potential for the SMRT platform to investigate the precise character of methylome, and significantly enhance our understanding of the function of DNA MTase.

INTRODUCTION

Tuberculosis (TB) has been the subject of a global health emergency, with more than 9 million new cases of active disease and nearly 1.5 million deaths annually (1). As the primary pathogen causing tuberculosis, *Mycobacterium tuberculosis* (MTB) belongs to the *Mycobacterium tuberculosis* complex (MTBC), all of whose members can cause tuberculosis in humans or other organisms. Besides *M. tuberculosis*, the MTBC group mainly consists of some genetically related *Mycobacterium* species (sharing more than 99% identity at the nucleotide level) including: *Mycobacterium bovis*; *M. bovis* BCG; *Mycobacterium africanum*; *Mycobacterium microti*; and *Mycobacterium canettii* (2). The MTBC members have different methods of host adaption and pathogenicity: *M. tuberculosis* and *M. africanum* mainly infect humans, whereas *M. bovis* now rarely causes disease in humans and mainly infects animals such as cattle. *M. microti* has been reported to mainly infect animals such as voles and also rarely infects humans, while the at-

*To whom correspondence should be addressed. Tel: +86 10 8409 7476; Fax: +86 10 8409 7845; Email: chenfei@big.ac.cn
Correspondence may also be addressed to Prof. Hairong Huang. Tel: +86 10 89509359; Fax: +86 10 89509359; Email: hairong.huangcn@gmail.com
†These authors contributed equally to this paper as the first authors.

tenuated *M. bovis* BCG strain is derived from a virulent *M. bovis* strain, and is well known as a vaccine against tuberculosis.

Among MTBC clinical isolates, the main human-adapted strain lineages are classified into seven groups (Lineage 1–7, L1–7) according to their geographic distribution: Indo-Oceanic (L1); East Asian (L2); East African-Indian (L3); Euro-American (L4); West African 1 (L5); West African 2 (L6); and Ethiopian (L7). L1–L4 comprise the majority of human-adapted strains which are responsible for global human TB cases, while L5 and L6 are restricted to West Africa, and L7 is only associated with those cases surfacing in the Horn of Africa (3). The animal-adapted strains are regarded as lineage 8 (L8) (4). Alternatively, based on the presence or absence of a *M. tuberculosis* specific deletion (TbD1), MTBC strains can also be classified as ancient strains (L1; L5–L7) and modern strains (L2–L4) (4).

There are some genomic studies concerning MTBC strains including DNA methylation, which have been reported in recent years (5–8). However, due to their high GC content (∼65%), and the PE/PPE multigene families (accounting for ∼10% of the genome) having repetitive sequences of PE/PPE regions (7), only 48 MTBC genomes have been completed so far, while most of the others (more than 1500) are only draft assemblies (http://www.ncbi.nlm.nih.gov/genome/genomes). Next-generation sequencing (NGS) platforms, such as the Illumina HiSeq platform, have had technical difficulties in finishing whole genomes with a high GC content and large numbers of repetitive sequences (9). Meanwhile, DNA methylation in MTBC genomes was also studied using various methods and technologies, including liquid chromatography-coupled tandem mass spectrometry (LC-MS/MS), restriction digest susceptibility and Sanger sequencing (8,10). This line of research revealed the existence of N6-methyladenine (^{m6}A) and 5-methylcytosine (^{m5}C) within MTBC genomes. Two DNA methyltransferases (MTases), MamA and HsdM, which are responsible for ^{m6}A modification, were discovered, of which MamA MTase has been well characterized. Its targeted methylated sequence motif (CTCCAG) was also identified. Interestingly, the function of MamA appears to be different from that of most MTases in prokaryotes, whose primary function of DNA methylation is DNA self-recognition via restriction-modification (R-M) systems that protect the organism against invading DNA (11). Instead, MamA appears to operate more like an 'orphan' MTase, such as Dam in *Escherichia coli*, because so far the cognate restriction enzyme has not been discovered in MTB genomes (12). It has been reported that the 'orphan' MTases may play an important role in chromosome stability, mismatch repair, replication, etc. (12,13). MamA has also been proved to be capable of influencing gene expression and fitness during hypoxia (13).

Single-molecule real-time (SMRT) sequencing, recently developed by Pacific Biosciences, can achieve unbiased GC coverage with extraordinarily long reads (up to 20 kb), therefore it is suitable for use in sequencing the genomes with high GC content and large numbers of repetitive regions like MTBC strains. Furthermore, SMRT sequencing allows direct genome-wide detection of diverse modified bases (^{m6}A, ^{m4}C, ^{m5}C, phosphorothioate modification, etc.) by monitoring the kinetic variations (KV) of the single base (14), in which ^{m6}A provides the strongest signals. Remarkably, SMRT sequencing provides a novel strategy for characterizing the 'precision methylome' by detecting the modification status of each site and even each sequence-read (15,16). This method enables the MTase activity to be intricately characterized at a whole-genome scale. In previous years, many bacterial methylomes (*E.coli*, *Mycoplasma genitalium*, etc.) have been determined using SMRT sequencing technology (17–20). However, until now, the panoramic analysis of MTBC's genomic methylation has not been reported. We have reason to believe that methylome characterization of MTBC strains will help us to completely understand the genomic function.

In this study, we completed the genomes of 12 MTBC strains and analyzed their respective methylomes using SMRT sequencing technology. The results revealed that there were three ^{m6}A motifs and three corresponding ^{m6}A DNA MTase genes encoded within the MTBC genomes, including the reported *mamA* (Rv3263), the *hsdM* (Rv2756c) (8) and a newly discovered MTase gene (Rv2024c), which we termed *mamB* (Mycobacterial adenine methyltransferase B). We report, to our knowledge for the first time, verification of the methylated motifs and functions for both the HsdM and MamB MTases. Interestingly, we found that the activities of the three MTases varied within different lineages, probably due to either mutations or deletions inactivating the corresponding MTases. Furthermore, based on the information obtained from SMRT sequencing, we investigated the methylation status of each modified site and even each modified sequence-read. By determining the 'the methylated-motif-site ratio' and 'the methylated-read ratio', we were able to discover the 'precision methylome' of the MTBC strains, which enabled an intricate analysis of MTase activity on the scale of the whole-genome.

MATERIALS AND METHODS

Bacterial strains and culture conditions

All MTBC strains used in this study are listed in Supplementary Table S1. The MTBC strains were grown in Lowenstein–Jenden media or Middlebrook 7H10 supplemented with 10% OADC (Oleic Albumin Dextrose Catalase, Becton Dickinson), glycerol and 0.05% Tween 80.

MIRU-VNTR genotyping

We used the VNTR-15 scheme as described in MIRU-VNTR*plus* (http://www.miru-vntrplus.org/) for genotyping, comprising the following markers: Mtub04; ETRC; MIRU04; MIRU40; MIRU10; MIRU16; Mtub21; QUB11b; ETRA; Mtub30; MIRU26; MIRU31; Mtub39; QUB26; and QUB4156. Each MIRU-VNTR locus was amplified individually, and electrophoresis of products on agarose gels was carried out as described previously (21). The copy number at each locus was calculated with BioNumerics software.

Large sequence polymorphism (LSP) determination of deletions in pks15/1

A large sequence polymorphism (LSP) (pks15/1) was used to differentiate the clinical TB strains by using the primers in Supplementary Table S2. Deletion of a 7-bp region in the polyketide synthase gene pks15/1 is present in the Euro-American lineage of *M. tuberculosis* (22).

SMRT sequencing

Genomic DNA from the MTBC strains was extracted using TIANamp Bacteria Genomic DNA Kit (Tiangen Biotech Co. Ltd., Beijing, China). Pacific Biosciences RSII DNA sequencing system (Pacific Biosciences, Menlo Park, CA, USA) was used as the sequencing platform. A 10-kb SMRTbell library was prepared from sheared genomic DNA (>5 μg) using a 10-kb template library preparation workflow according to the manufacturer's recommendation, with an additional bead clean-up before primer annealing. The library was bound with P4 polymerase and complexes were loaded on to version V3 SMRT cells. These were sequenced using 165 min movies. Two SMRT cells were used for each strain, yielding output data with average genome coverage of ∼100X.

Bioinformatic analyses of SMRT sequencing data

De novo assembly of the insert reads was performed with the Hierarchical Genome Assembly Process (HGAP.3) algorithm in SMRT Portal (version 2.2.0). Gap closing was finished by PBJelly (23). Circularization was achieved by manual comparison and removal of a region of overlap, and the final genome was confirmed by remapping of sequence data. rRNA and tRNA predictions were performed using RNAmmer (24) and tRNAscan-SE (25), respectively. Genes were predicted using Prodigal version 2.60 (26). We first assigned annotation by comparison with the well annotated H37Rv (GenBank accession NC_000962) and the NCBI non-redundant (NR) database was used to further assign these annotations for all the sequences with no hits. The average nucleotide identity with H37Rv (NC_000962) was calculated through ANI on EzGenome (http://www.ezbiocloud.net/ezgenome/ani). Promoter regions were analyzed using Neural Network Promoter Prediction (http://www.fruitfly.org/seq_tools/promoter.html) and PePPER (http://pepper.molgenrug.nl/index.php/prokaryote-promoters).

Genome-wide detection of base modification and the sequence motifs were performed using the standard settings (QV of modified motifs are more than 30) in the 'RS_Modification_and_Motif_Analysis.1' protocol included in SMRT Portal version 2.2.0. 'Motif score' = (number of detections matching motif) / (number of genome sites matching motif) × (sum of log-p value of detections matching motif) = (fraction methylated) × (sum of log-p values of matches). SMRT Portal searches (close to maximize) through the space of all possible motifs, progressively testing longer motifs using a branch-and-bound search. The minimal term 'motif score' is 40 and 'fraction methylated' term must be less than 1. Furthermore, the MTase target sequence motifs were identified by selecting the top 1000 kinetic hits and subjecting a ±20 base window around the detected base to MEME-ChIP (27), and then compared with the predictions in REBASE (28).

Through measuring 'the methylated-motif-site ratio' and 'the methylated-read ratio', we were able to obtain the 'precision methylome' of the MTBC strains. 'The methylated-motif-site ratio' represented the percentage of modified motif sites within all the motif sites and was obtained automatically by SMRT Portal (RS_Modification_and_motif_Analysis protocol, Version 2.2.0). 'The methylated-motif-site ratio' = 'number of methylated-motif-sites' / 'number of motif-sites' × 100%.

'The methylated-read ratio' represented the percentage of the modified reads within the total number of reads (modified reads + unmodified reads) which were mapped to the site. It was also calculated automatically by 'RS_Modification_and_motif_Analysis' protocol (Version 2.2.0). 'The methylated- read ratio' = 'number of reads mapped to the sites which have a modified base' / 'number of all reads mapped to the sites' × 100%. Here, in 'modification.sff' file, the numerical result of 'methylated-read ratio' of the motifs are from 'frac = decimal'. 'Frac' represents the 'fraction' of reads mapped to the position which have a modified base (e.g. 'frac = 0.80' means 'methylated-read ratio' of the base is 80%).

Single nucleotide polymorphism (SNP) identification and Phylogenetic analysis

Single nucleotide polymorphisms (SNPs) were analyzed by MAUVE 2.3.1 (29) using the 12 genomes in the study and 22 published genomes (Supplementary Table S3). Paired-end reads from an *M. tuberculosis* Ethiopia strain were mapped to the H37Rv reference genome (GenBank accession NC_000962) using BWA 0.5.9 (30) and SNPs were called using SAMtools 0.1.19 (31). SNPs called in repetitive regions of the H37Rv genome, defined as exact repetitive sequences of ≥25 bp in length, identified using RepeatMasker (32), were excluded. The SNPs present in all strains were used to construct a phylogenetic tree on the basis of maximum likelihood using MEGA 6.06 (33). A strain of *Mycobacterium canettii* was included as an outlier.

^{m5}C bisulfite sequencing and data analysis

^{m5}C methylation was detected by bisulfite sequencing (34). Genomic DNA (100 ng) was used to construct the DNA library. They were sonicated into 100–500 bp fragments and then subjected to end repair, dA tailing, ligation and gel purification. Bisulfite conversion was performed using the EZ DNA methylation-Gold kit (Zymo Research). The library was sequenced by HiSeq 2000. Adapter and low-quality base trimming was performed by Trimmomatic version 0.32 (35) with default parameters. Filtered paired-end reads were mapped against the reference genomes by Bismark (version 0.12.2) (36).

MTase cloning

MamB and HsdM MTase genes (Rv2024c and Rv2756c) were amplified from the genomic DNA of *M. bovis* 30 and

cloned into the plasmid pRRS as described previously (37). The gene-specific oligonucleotide primers used for PCR are described in Supplementary Table S2. Restriction sites diagnostic for the predicted methylation pattern were incorporated into the 3′-end oligonucleotides. The presence or absence of specific methylation was determined by digesting the constructs with appropriate restriction enzymes and sequencing the plasmid with the MTase gene by PacBio RS II. The host strain used for cloning was *E. coli* ER2796 (38). *E. coli* strains were cultured in LB broth or on LB plates supplemented with ampicillin (final concentration 100 g/ml) as required. Various truncated Type I MTases were established to validate the function of both the modification subunit (HsdM) and the specificity subunit (HsdS).

Restriction digests

Plasmid DNA was prepared with a Qiagen Miniprep kit and then cleaved with the appropriate diagnostic restriction endonucleases (BclI and Eco47III, respectively (New England Biolabs)) for assaying the methylation state for HsdM and MamB MTase. Digestion reaction were performed for 2 h at 37°C and run on 1% agarose gels.

Evolution of the DNA MTases

A total of 2775 complete genomes (1493 species) of bacteria from the NCBI RefSeq database (2014-10-23) were downloaded, and the sequences of three MTases were searched in the database by BLAST+ 2.2.30 (39). We selected all species which had a protein identity of above 30% with the three respective DNA MTases and retained the best result for each species. The multiple sequence alignment of all homologous proteins was generated using ClustalW (40) embedded in MEGA 6.06 (33). Based on the protein identity results, we built a probable phylogenetic tree of the three DNA MTases incorporating the maximum likelihood method using MEGA 6.06. iTOL was used for display and annotation of phylogenetic trees (41).

RESULTS

General bioinformatic analysis for 12 MTBC strains

Twelve MTBC strains, including six reference strains and six clinical isolates, were sequenced by SMRT sequencing technology (14). All strains were *de novo* sequenced, including *M. microti* which had not been sequenced before. The average of the sequencing coverage was ∼100X. These new genomes were all completed with the use of SMRT Portal and PBJelly software (Supplementary Figure S1). All the genome data have been deposited in NCBI with the GenBank accession numbers CP010329–CP010340 for *M. tuberculosis* (Mtb) F1, Mtb F28, *M. bovis* BCG 26, *M. bovis* 30, *M.microti* 12, *M. africanum* 25, Mtb 2242, Mtb 2279, Mtb22115, Mtb37004, Mtb 22103 and Mtb 26105, respectively. The bioinformatic analysis provided the general genome information (Table 1 and Supplementary Table S4), including GC% content (∼65.6%); the size of the respective genomes (∼4.34–4.43 Mb); and the number of predicted protein-coding genes (∼4000–5000). As shown in Supplementary Figure S1 and Table S4, these predicted genes were generally distributed evenly across the plus and minus strands, which accounted for the vast majority of the whole genome as consistent with other prokaryotes (42).

When compared with the *M. tuberculosis* H37Rv reference genome (NC_000962) (7), our 12 MTBC genomes showed more than 99.70% identity at the nucleotide level when using ANI on EzGenome (http://www.ezbiocloud.net/ezgenome/ani). The genome structure of all the MTBC strains was then explored by multiple genome alignment using progressive MAUVE (29). No extensive translocations, duplications or inversions were found in the genomes except for strain Mtb 2242, which contained a 1.8 Mb inverted fragment with the same transposase gene on both ends (Supplementary Figure S2). We investigated further by designing different primer sets in order to amplify the junction regions by PCR. These PCR results confirmed the presence of the 1.8 Mb inversion (data not shown).

To implement the phylogenetic analysis, a total of 30 598 SNPs were discovered in the non-repetitive regions of the MTBC genomes and then used to construct a genome-wide maximum-likelihood phylogeny (Figure 1, Supplementary Table S5). Our tree topology comprised seven main lineages (L1–L7) and one animal-adapted lineage (L8), which matched satisfactorily with the published trees (6,43). Of the six clinical isolates, two strains (Mtb 2242 and Mtb 2279) belonged to L2 (Beijing type); three strains (Mtb 22115, Mtb 37004 and Mtb 22103) belonged to L4 (Euro-American lineage); and the remaining strain (Mtb 26105)

Table 1. The general genome information of 12 MTBC strains

Strain No.	Species	ATCC No./Lineage	Average read size (kb)	Sequencing coverage (x)	Completed genome size (Mb)	GC content (%)	Gene number	Average gene size (bp)
F1	*M. tuberculosis* H37Rv	27294/L4	4.39	120	4.43	65.61	4320	927
F28	*M. tuberculosis* H37Ra	25177/L4	4.03	94	4.42	65.60	4179	959
30	*M. bovis*	19210/L8	4.39	94	4.34	65.60	4198	932
26	*M. bovis* BCG	35735/L8	5.37	99	4.35	65.61	4780	809
12	*M.microti*	19422/L8	6.81	128	4.37	65.63	4323	910
25	*M. africanum*	35711/L6	5.30	95	4.39	65.58	4801	813
2242	*M. tuberculosis*	L2	4.90	89	4,42	65.60	4467	888
2279	*M. tuberculosis*	L2	5.13	97	4.41	65.59	4601	858
22115	*M. tuberculosis*	L4	5.74	123	4.40	65.57	4213	946
37004	*M. tuberculosis*	L4	6.66	125	4.42	65.60	4231	943
22103	*M. tuberculosis*	L4	4.37	101	4.40	65.61	4186	952
26105	*M. tuberculosis*	L3	4.97	101	4.43	65.63	4200	952

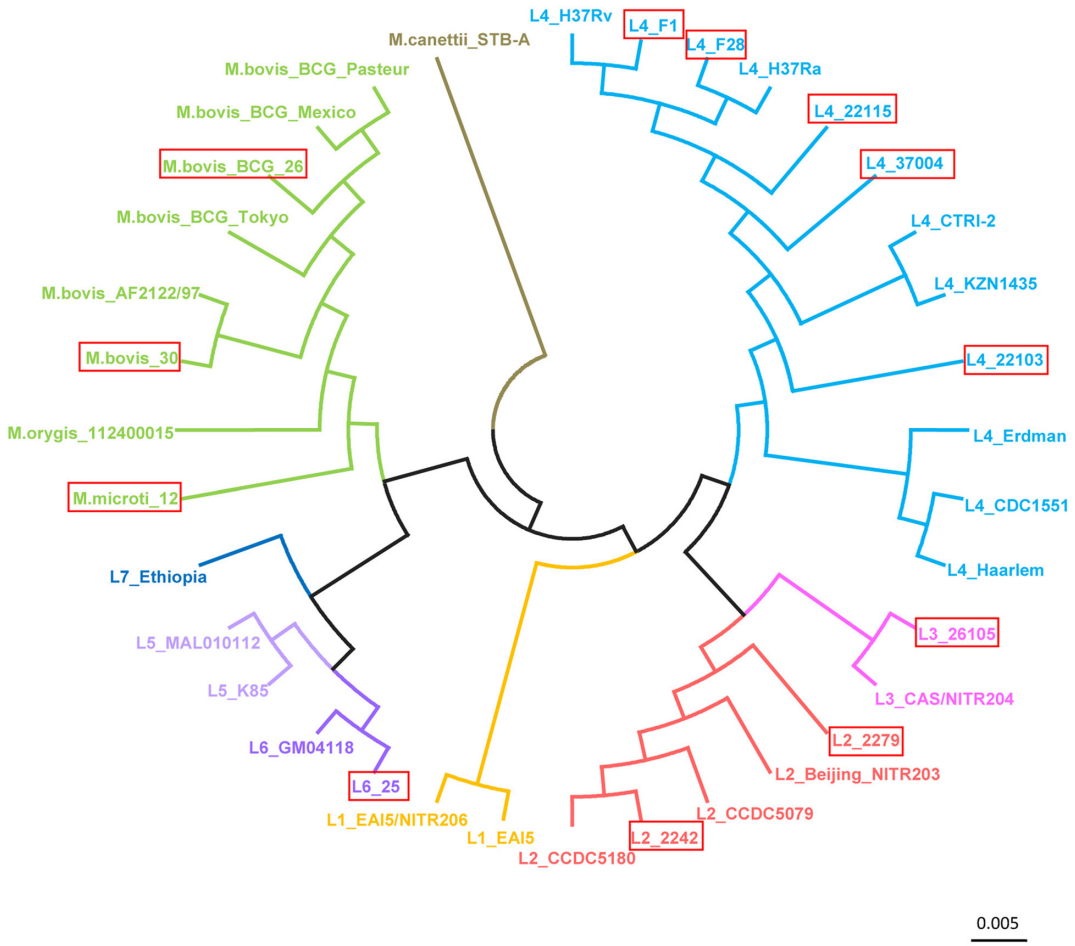

Figure 1. Phylogenetic analysis of Mycobacterium tuberculosis complex (MTBC) strains. Blue indicates that strains belong to lineage 4 (L4, Euro-American lineage). Pink indicates that strains belong to L3 (East African-Indian lineage). Red indicates that strains belong to L2 (East Asian lineage). Orange indicates that strains belong to L1 (Indo-Oceanic lineage). Purple indicates that strains belong to L5 and L6 (West African 1 and West African 2 lineage). Dark blue indicates that strains belong to L7 (Ethiopia lineage). Green indicates that strains belong to animal-adapted lineage (L8). The strains in red boxes represent the 12 MTBC genomes we finished here.

belonged to L3 (East African-Indian lineage). Of the reference strains, *M. africanum* 25 belonged to L6 (West African 2) while, as expected, the other reference strains belonged to the corresponding lineages: Mtb F1 and Mtb F28 to L4 (Euro-American lineage); and *M. bovis* 30, *M. bovis* BCG 26, and *M. microti* 12 to the animal-adapted lineage (L8).

DNA methylome analysis of 12 MTBC strains using SMRT sequencing

The genome-wide distribution of methylated bases in 12 MTBC strains was determined using SMRT sequencing technology. Through bioinformatic analysis of the SMRT sequenced data, N6-methyladenine (^{m6}A) could be detected with a high level of accuracy. In addition to ^{m6}A, a significant number of bases were reported as 'modified bases'. Among these suspected sites, almost all of the reported bases failed to show any obvious methylated patterns after analysis with the SMRT Portal analysis platform. The reported modifications were subsequently verified by whole genome amplification (WGA) technology as previously described, this being based on the multiple displacement amplification (MDA) method (44). This WGA testing revealed that almost all of the modifications which did not have specific patterns were in fact false positives. We also observed that the IPD values (average 7.28) of ^{m6}A were usually higher than those of the other modifications (average 2.13). Incidentally, no ^{m5}C modification was found in the genomes of the twelve MTBC strains when using bisulfite sequencing, this being consistent with some previous reports (8).

Three ^{m6}A sequence motifs were detected in the 12 MTBC strains by SMRT Portal analysis (Table 2). In addition to the previously identified methylated sequence motif CTCCAG (5′ to 3′ direction; the detected methylated base was indicated A, and T represented the thymine pairing with the methylated adenine on the complementary strand) (8), 2 adenine methylation motifs were also identified in the 12 MTBC strains: an asymmetric motif CACGCAG and a characteristic Type I motif GATN$_4$RTAC (Supplementary Figure S1 and Table 2). This concurred with the predictions presented on the REBASE website (http://tools.neb.com/~vincze/genomes/index.php?page = M). The methylation of the motifs varied in the 12 MTBC strains, with all of the 3 methylated motifs being discovered in 4 of the ancient strains (*M. africanum*; *M. microti*; *M. bovis* BCG; and *M.*

Table 2. Comparison of genome-wide methylation patterns among 12 MTBC strains

Strain No.	Sequence motif[b]					
	CTCC**A**G		CACGC**A**G		G**A**TN$_4$R**T**AC	
	No. of motif[a]	% modified motif	No. of motif	% modified motif	No. of motif	% modified motif
Mtb F1	3902	99.67	825	0	724	0
Mtb F28	3904	99.72	821	0	730	0
M. bovis 30	3828	99.76	803	100	686	87.03
M. bovis BCG 26	3842	99.53	806	100	678	77.73
M.microti 12	3860	99.87	817	100	698	97.56
M. africanum 25	3872	99.5	813	100	706	90.65
Mtb 2242	3918	0	829	100	730	96.85
Mtb 2279	3902	0	824	100	722	97.23
Mtb 22115	3890	99.59	817	100	716	0
Mtb 37004	3894	96.92	820	100	714	0
Mtb 22103	3876	99.66	815	100	700	96
Mtb 26105	3908	99.56	824	100	724	0

[a]The number of motifs includes ones on the plus and minus strands.
[b]The methylated nucleotide in the motif is shown as bold and underlined letter. The underlined letter represents the thymine pairing with the methylated adenine on the complementary strand.

bovis). However, this was not the case with the modern strains. We found two methylated motifs (G**A**TN$_4$R**T**AC and CACGC**A**G) in the L2 clinical isolates (Mtb 2242 and Mtb 2279), two methylated motifs (CTCC**A**G and CACGC**A**G) in the L3 clinical isolate Mtb 26105, while for the L4 lineage strains, it became more complicated. The L4 reference strains, *M. tuberculosis* H37Rv and *M. tuberculosis* H37Ra, contained only one methylated motif CTCC**A**G; two L4 clinical isolates (Mtb 22115 and Mtb 37004) contained two methylated motifs (CTCC**A**G and CACGC**A**G); and another L4 clinical isolate, Mtb 22103, contained all of the three modified sequence motifs which were found in the ancient strains (Table 2). The unmethylated motifs in these genomes could be ascribed to the functional inactivation of the three related MTase genes which contained some missense mutations and partial deletions (discussed in detail in the next section).

The three motifs were then plotted as different colored tracks on the Circos plots to exhibit their distribution across the whole genomes of the 12 MTBC strains. This showed that they were distributed randomly across these genomes (Supplementary Figures S1 and S3). We then analyzed the distribution of the motifs separately within gene regions (GRs) and intergenic regions (IGRs). The motif G**A**TN$_4$R**T**AC was identified as being preferentially located within IGRs when compared with the other two motifs: about 12–13% of the G**A**TN$_4$R**T**AC motifs were located in IGRs, whereas less than 10% of the other motifs CTCC**A**G and CACGC**A**G were within IGRs (Supplementary Table S6 and Figure S4). We subsequently calculated the GC content of IGRs and GRs, respectively. The average GC content of IGRs and GRs is 63.4% and 65.8%, which may account for the difference in distribution of the motif G**A**TN$_4$R**T**AC (12–13%) and other two motifs (10%). Further analysis of the motifs occurring in IGRs revealed a relative enrichment of the motif G**A**TN$_4$R**T**AC at −70 to −80 bp from the start codon (Supplementary Figure S5A). A promoter was also predicted to be located in this region by using Neural Network Promoter Prediction (http://www.fruitfly.org/seq_tools/promoter.html) and PePPER (http://pepper.molgenrug.nl/index.php/prokaryote-promoters). This proximity to the promoters suggested that the methylated sequence motif might be involved in regulating promoter activity. Incidentally, relative enrichments of motifs CTCC**A**G and CACGC**A**G were found at −10 bp to −20 bp and −30 bp to −40 bp from the start codon, respectively (Supplementary Figure S5B and S5C).

We then proceeded to investigate the COG functional category of genes containing three sequence motifs (Supplementary Figure S6). Among the COG categories, only motif G**A**TN$_4$R**T**AC displayed a significant enrichment in COG category L (replication, recombination and repair) in two L2 clinical isolates (Mtb 2242 and Mtb 2279). This is because these two L2 clinical isolates contained higher copies of IS6110 with the sequence of motif G**A**TN$_4$R**T**AC.

Analysis of the unmethylated motif

SMRT sequencing revealed that not all modified motif sites were detectable as being modified (17). The results in Table 3 show that there were some G**A**TN$_4$R**T**AC and CTCC**A**G sites which were detected as always being unmethylated within the MTBC genomes. Amongst these sites, most were detected as being unmethylated on both strands, but some sites were shown as being hemi-methylated (methylated on only one strand). The sequence coverage (∼100X) was sufficient to support these findings (17). From a total of about 4000 CTCC**A**G sites within the MTBC genomes, fewer than 20 (5–20 loci, ≤0.5%) were identified as always being unmethylated. The only exception was the Mtb 37004 genome which contained 122 unmethylated sites (about 3%). With regard to the G**A**TN$_4$R**T**AC motif (∼700 loci), the genomes of three ancient strains (*M. africanum* 25; *M. bovis* BCG 26; and *M. bovis* 30) also had more unmethylated sites (66–151 loci, 9.4–22.3%).

To investigate the distribution of unmethylated sites within the 12 MTBC strains, we analyzed their positions in both GRs and IGRs. As shown in Table 3, a total of

Table 3. The unmethylated sites among 12 MTBC strains

Strain No.	Unmethylated sites (G**A**TN$_4$R**T**AC)				Unmethylated sites (C**T**CC**A**G)			
	Total	GR	IGR	% in IGR	Total	GR	IGR	% in IGR
Mtb F1	/	/	/	/	11(5)	10(4)	1(1)	9.09
Mtb F28	/	/	/	/	11(3)	10(2)	1(1)	9.09
M. bovis 30	89(11)[a]	73(9)	16(2)	17.98	9(3)	9(3)	0	0.00
M. bovis BCG 26	151(21)	124(20)	27(1)	17.88	18(12)	15(11)	3(1)	16.67
M.microti 12	17(5)	9(3)	8(2)	47.06	5(3)	4(2)	1(1)	20.00
M. africanum 25	66(14)	54(10)	12(4)	18.18	20(12)	18(10)	2(2)	10.00
Mtb 2242	23(3)	14(2)	9(1)	39.13	/	/	/	/
Mtb 2279	20(4)	12(4)	8(0)	40.00	/	/	/	/
Mtb 22115	/	/	/	/	16(8)	15(7)	1(1)	6.25
Mtb 37004	/	/	/	/	122(62)	113(59)	9(3)	7.38
Mtb 22103	28(6)	22(6)	6(0)	21.43	13(5)	13(5)	0(0)	0.00
Mtb 26105	/	/	/	/	17(5)	14(4)	3(1)	17.65
Average				22.99				9.09

GR: Gene Region; IGR: Intergenic Region.
[a]The number in the parentheses indicates the number of hemi-methylated sites.

~23% of unmethylated G**A**TN$_4$R**T**AC sites were located within the IGRs in seven MTBC strains, which was ~10% above the proportion of the other sites in IGRs (~12%, Supplementary Table S6). In *M. microti* 12, Mtb 2242 and Mtb 2279, the proportions were even as high as 40–50%. To systematically investigate this location bias of the unmethylated G**A**TN$_4$R**T**AC sites within IGRs, the distances between the unmethylated sites and the start codon were determined (Supplementary Figure S7). In comparison with most of the MTBC strains, the unmethylated G**A**TN$_4$R**T**AC sites were located within 50 bp upstream from the start codon. Since most of the promoters in the 12 MTBC strains were predicted to be located 70–80 bp upstream from the start codon, these sites might always remain unmethylated so as to ensure that the related gene is transcribed smoothly. With regard to the unmethylated C**T**CC**A**G sites, they had approximately the same distribution proportion within IGRs (~9.1%, Table 3) as the motif did (6.1–8.2%, Supplementary Table S6).

Exploration of the precision methylome by determining 'the methylated-motif-site ratio' and 'the methylated-read ratio'

In order to investigate the precision methylome across all of the 12 MTBC genomes, we determined 'the methylated-motif-site ratio' by calculating the percentage of modified motif sites within all the motif sites (modified motif sites + unmodified motif sites) (Table 2). For the C**T**CC**A**G motifs, except Mtb 37004 (about 97%), the methylated-motif-site ratios for the MTBC strains were greater than 99.5%. However, this was not the case for the G**A**TN$_4$R**T**AC motifs: the methylated motif site ratios (about 77.7–90.7%) for the three ancient strains (*M. africanum* 25; *M. bovis* BCG 26; and *M. bovis* 30) were noticeably lower than those for the other strains (about 97%). In addition, after calculating the methylated-motif-site ratio for G**A**TN$_4$R**T**AC with the permutations of the four Ns, it could be seen that no nucleotide selection preference of the N sites was found to be responsible for the differences between these strains (data not shown). With regard to the CACGC**A**G motifs, all the sites (100%) were determined to be fully methylated for the tested strains with active MTases.

Looking at each methylated motif site, not all of the sequence-reads covering the site were methylated (15,17). To investigate the DNA methylation heterogeneity (or partial methylation), we analyzed 'the methylated-read ratio' of three motifs in the MTBC strains by calculating the percentage of the modified reads within the total number of reads (modified reads + unmodified reads) which were mapped to the site. This provided us with a more precise way of characterizing the methylome. Figure 2A showed that

Table 4. The average methylated read ratio of 12 MTBC strains

Stain No.	C**T**CC**A**G	CACGC**A**G	G**A**TN$_4$R**T**AC
Mtb F1	96.2	/	/
Mtb F28	96.17	/	/
M. bovis 30	94.5	97.2	88.81
M. bovis BCG 26	95.7	97.29	82.43
M.microti 12	95.86	97.6	95.51
M. africanum 25	94.28	97.02	90.41
Mtb 2242	/	97.99	95.73
Mtb 2279	/	97.1	94.05
Mtb 22115	95.69	98.52	/
Mtb 37004	82.06	98.38	/
Mtb 22103	94.47	97.7	93.82
Mtb 26105	95.88	98.49	/

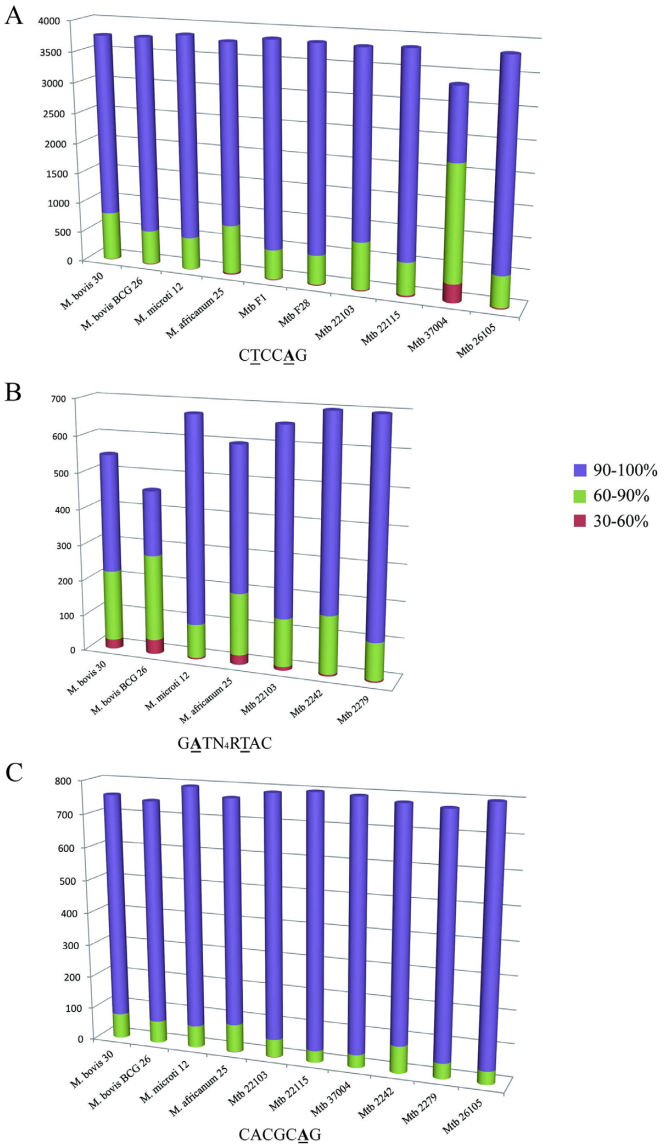

Figure 2. Distribution of the methylated read ratio of three motifs in the MTBC stains. The horizontal axis shows the strain name. The vertical axis shows the number of motifs with diverse methylated read ratio (30–60%, 60–90%, 90–100%). The methylated read ratio indicates the percentage between the reads containing the methylated base and the total reads mapped to the site. For example, as for one methylated site, if its methylated read ratio is 60% and there are 100 reads covering (mapped to) this methylated site, this means that 60 reads contain the methylated base and the other 40 reads have no methylation. (**A**) Distribution of the methylated read ratio of CTCCAG motif in the MTBC stains. (**B**) Distribution of the methylated read ratio of GATN$_4$RTAC motif in the MTBC stains. (**C**) Distribution of the methylated read ratio of CACGCAG motif in the MTBC stains.

most of the CTCCAG sites (78.25–87.11%) were almost fully methylated (methylated-read ratio: 90–100%) in the MTBC strains with the exception of Mtb 37004, with the average methylated-read ratio (82.06%) of Mtb 37004 being significantly lower than that for the other strains (average 95.40%) (Table 4). Here, only 34.58% of the CTCCAG sites in Mtb 37004 showed a high-level methylated-read ratio (90–100%); 56.42% displayed a mid-level methylated-read ratio (60–90%); and 9% exhibited a low-level methylated-read ratio (30–60%) (Figure 2A). In contrast, the average methylated-read ratios of the GATN$_4$RTAC sites in the three ancient strains (*M. africanum* 25, 90.41%; *M. bovis* BCG 26, 82.43%; and *M. bovis* 30, 88.81%) were lower than the other four MTBC strains (93.82–95.73%) (Table 4 and Figure 2B), while, as shown in Figure 2C, most of the CACGCAG sites (88.89–95.48%) showed a high level methylated-read ratio (90–100%) in the MTBC strains. Interestingly, we observed that there were very few motifs with a 100% methylated-read ratio, which was probably because the methylated motifs underwent passive demethylation (45) during replication before they had time to be methylated again.

DNA methyltransferases in MTBC strains

In the search for genes which were homologous to known DNA MTases in the REBASE database (28), we identified three homologous genes which were responsible for the three methylated motifs in the MTBC genomes (Figure 3). Apart from the reported *mamA* (8), the other two genes were predicted to separately encode a Type I MTase and a Type IIG MTase. Type I MTase (M) gene is generally located between its cognate restriction enzyme gene (R) and a specificity subunit encoding gene (S). The three parts (S–M–R) together function as a Type I system which recognizes a bipartite motif comprising two short sequences (3–5 nt) separated by 5–8 non-specific nucleotides (46). The predicted Type IIG MTase was encoded by the Rv2024c gene which we termed *mamB*. In general, Type IIG MTases only methylate one strand of the asymmetric recognition sequences (47).

In order to determine whether the two predicted MTases were responsible for the two newly detected methylated motifs (GATN$_4$RTAC and CACGCAG), we implemented the restriction digestion and SMRT sequencing of plasmids containing cloned MTase genes in methyltransferase-free *E. coli* ER2796 (38). It has been shown previously that the Type I MTase encoded by Rv2756c gene, which was named as HsdM, is responsible for ^{m6}A modification (8), but its methylated motif has not been characterized. The genome sequence analysis showed that the core of the HsdM Type I MTase consisted of three subunits: specificity subunit 1 (designated as S1); MTase subunit (designated as M); and specificity subunit 2 (designated as S2). Four genes encoding two pairs of antitoxin and RNase (designated as A1, R1, A2 and R2, respectively) were present between M and S2 (Supplementary Figure S8A1). However, the cognate restriction enzyme was not found in the MTBC strains, suggesting that it might be an orphan MTase. An MTase was defined as an orphan MTase if we were unable to detect a restriction endonuclease with the same target site as the MTase in the proximity of the MTase gene (48). To determine the function of the four parts (S2-R2A2R1A1-M-S1) of HsdM Type I MTase, different subunit-combinations were cloned and tested for their resistance to BclI cleavage (the methylation-sensitive BclI restriction site (5′-TGATCA-3′) overlaps the HsdM methylated motif site (GATN$_4$RTAC)) (Supplementary Figure S8A1). Regarding the full length of the sequence (S2-R2A2R1A1-M-S1), S2-M-S1 and S2-M, the electrophoresis results showed an almost com-

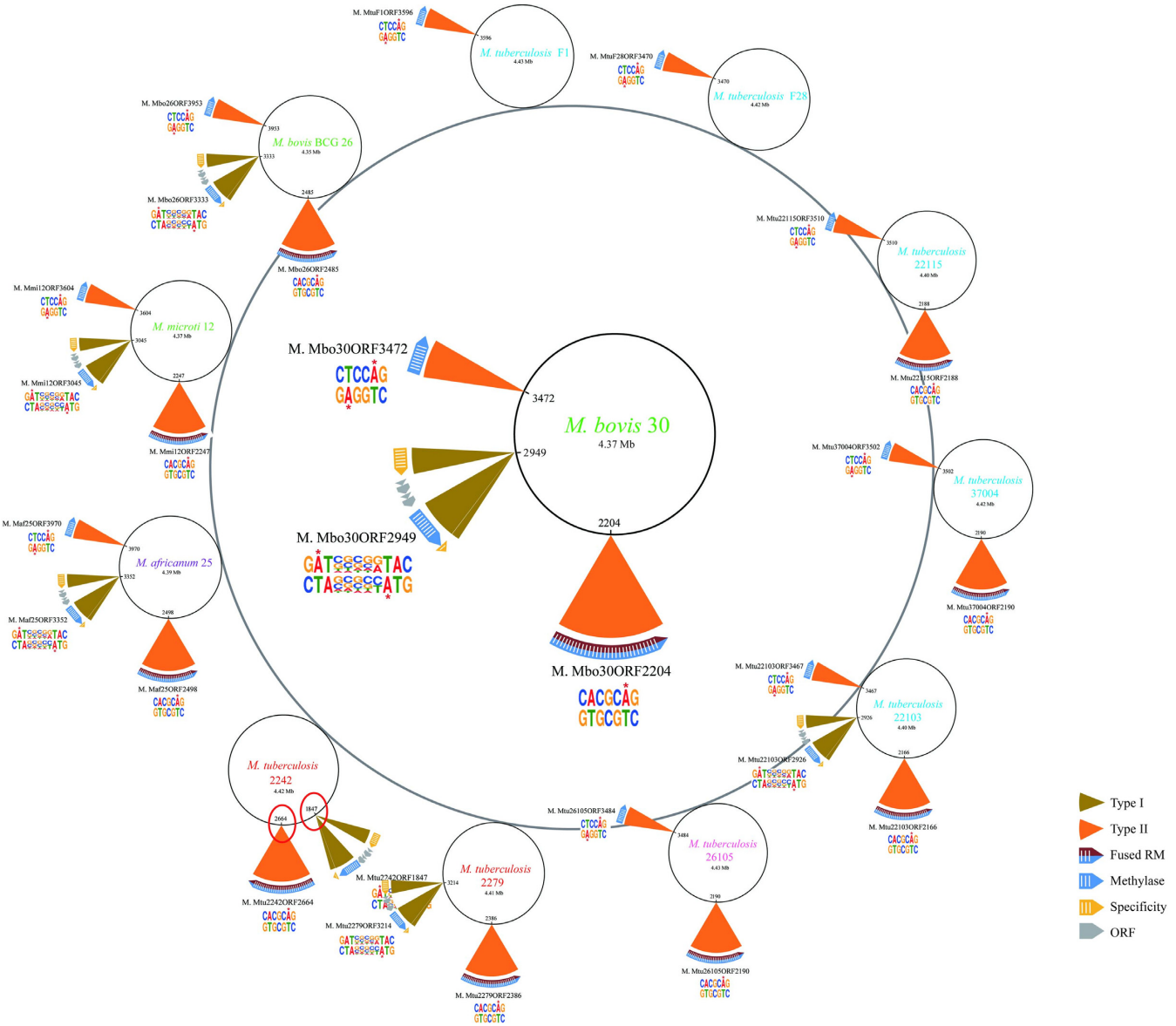

Figure 3. Three MTase genes and corresponding methylated sequence motifs in 12 MTBC strains. *M. bovis* 30 strain is located in the central position and the 11 other MTBC strains are located around it. Only active MTase genes were shown in the figure. The red circles mark the order change of the two MTase genes due to a large-scale inversion (about 1.8 Mbp) in Mtb 2242.

plete resistance to BclI cleavage (Supplementary Figure S8A2: lanes 5, 8 and 11). However, no resistance to BclI cleavage was observed at the GATN$_4$RTAC motif site if only the M-S1 genes were present (Supplementary Figure S8A2: lane 14). As expected, the cleavage results were confirmed by SMRT sequencing of the plasmids with different subunit-combinatorial *hsdM*-expressing clones (Supplementary Figure S8A3). Therefore, we concluded that HsdM was able to act on the motif GATN$_4$RTAC, and that during this process the S2 subunit was essential for MTase activity while the S1 subunit and the four gene products (between the S2 and M subunits) appeared to be dispensable by comparison. The reason why S1 subunit is dispensable may be because the target recognition domain (TRD) for S1 subunit was predicted not exist using the Interpro database (http://www.ebi.ac.uk/Tools/pfa/iprscan5/). Using the same method, we also confirmed that the CACGCAG motif was recognized and methylated by the MamB MTase (Supplementary Figure S8B1), while, similar to the situation with most organisms with Type IIG MTases, only hemi-methylation was detected at the asymmetric recognition site (Supplementary Figure S8B2) (47).

Based on the methylated motifs detected by SMRT sequencing, all the active MTases of the 12 MTBC strains were determined (Figure 3). The relative location of the *mamB* and *hsdM* in the Mtb 2242 genome was different from the others due to the large genomic inversion region (1.8 Mb) (Supplementary Figure S2). Subsequently, through multiple sequence alignment for all the MTBC MTase sequences, some mutations and deletions were dis-

Table 5. Distribution of SNPs/Indels in three MTBC MTase genes among 12 MTBC strains

A

Strain No.	HsdS2	HsdM					Activity
Mtb F1	/	/	/	C917T(P306L)	/	/	Inactive
Mtb F28	/	/	/	C917T(P306L)	/	/	Inactive
M. bovis 30	/	C279G	/	/	/	/	Active
M. bovis BCG 26	/	C279G	/	/	/	/	Active
M.microti 12	/	C279G	/	/	/	/	Active
M. africanum 25	/	C279G	/	/	G1374T(N458K)	A1442C(E481A)	Active
Mtb 2242	/	/	/	/	/	/	Active
Mtb 2279	/	/	/	/	/	/	Active
Mtb 22115	/	/	/	C917T(P306L)	/	/	Inactive
Mtb 37004	/	/	/	C917T(P306L)	/	/	Inactive
Mtb 22103	/	/	/	/	/	/	Active
Mtb 26105	T356G(L119R)	/	G518A(G173D)	/	/	/	Inactive

B

Strain No.	MamA			Activity
Mtb F1	/	/	/	Active
Mtb F28	/	/	/	Active
M. bovis 30	/	/	/	Active
M. bovis BCG 26	/	/	/	Active
M.microti 12	/	/	/	Active
M. africanum 25	/	/	/	Active
Mtb 2242	/	/	A809C(E270A)	Inactive
Mtb 2279	G61A(D21N)	/	A809C(E270A)	Inactive
Mtb 22115	/	/	/	Active
Mtb 37004	/	G527A(G176D)	/	Active
Mtb 22103	/	/	/	Active
Mtb 26105	/	/	/	Active

C

Strain No.	MamB											Activity
Mtb F1	C139T(R47W)	/	G461A(G154D)	/	/	/	/	Truncated (1520-4821)				Inactive
Mtb F28	C139T(R47W)	/	G461A(G154D)	/	/	/	/	Truncated (1520-4821)				Inactive
M. bovis 30	/	/	/	/	/	/	C865T(R289C)	/	/	/	/	Active
M. bovis BCG 26	/	/	/	/	/	/	C865T(R289C)	/	/	/	/	Active
M.microti 12	/	/	/	/	A661C(I221L)	/	C865T(R289C)	G1269A	/	/	/	Active
M. africanum 25	/	/	/	/	/	/	C865T(R289C)	/	/	A1972G(T658A)	/	Active
Mtb 2242	/	/	/	/	/	C696T	/	/	/	/	/	Active
Mtb 2279	/	/	/	/	/	/	/	/	G1531A(V511M)	/	/	Active
Mtb 22115	/	G141A	/	G567A	/	/	/	/	/	/	C2448G	Active
Mtb 37004	/	/	/	/	/	/	/	/	/	/	C2448G	Active
Mtb 22103	/	/	/	/	/	/	/	/	/	/	C2448G	Active
Mtb 26105	/	/	/	/	/	/	/	/	/	/	/	Active

Synonymous mutations are marked as blue; Non-synonymous mutations are marked as red.

"/" indicates no mutations in corresponding strains.

Underlined mutations indicate the reported mutations inactivating the MTases.

covered (Table 5). Combining this with the analysis of the MTase activities, it is therefore reasonable to infer the impact of these mutations and deletions on the related MTase activities. On the one hand, we found that some missense mutations did not influence the MTase activity, as all the three MTases were active within the four ancient strains (*M. africanum* 25; *M. microti* 12; *M. bovis* BCG 26; and *M. bovis* 30) and one L4 clinical isolate (Mtb 22103). From this we could deduce that the missense mutations G1374T and A1442C in *M. africanum* 25 had no significant effect on the HsdM activity. Additionally, the missense mutations: A661C in *M. microti* 12; C865T in the four ancient strains (*M. africanum* 25; *M. microti* 12; *M. bovis* BCG 26; and *M. bovis* 30); and A1972G in *M. africanum* 25, did not affect the activity of MamB. Among these, G1374T was also reported in L6 (43). On the other hand, we could infer from Figure 3 and Table 5 that some missense mutations and deletions could disrupt the activities of MTases within MTBC strains. Here, the activity of HsdM was lost within two reference L4 strains (Mtb F1 and Mtb F28) and two L4 strains (Mtb 22115 and Mtb 37004) due to the missense mutation C917T (Pro306Leu) as detailed previously (8). From Table 5A, we could reasonably infer that the missense mutation G518A (Gly173Asp) and/or T356G (Leu119Arg) within the L3 clinical isolate (Mtb 26105) might lead to the inactivation of the HsdM. This missense mutation (G518A) was also reported for typing L3 (43). As detailed in the literature, we also found that MamA became inactive as a result of the reported missense mutation A809C (Glu270Ala) (Table 5B) (8). With regard to MamB, it was the deletion in RvD1 (H37Rv related deletion) that caused MTase activity to be defective within two reference L4 strains (Mtb F1 and Mtb F28) (Table 5C). This is the first time that all of the above missense mutations have been reported to be associated with MTase activity within MTBC strains, with the exception of C917T in the *hsdM* and A809C in *mamA* (8).

These two important mutations that inactivate the MTases were investigated further by using the 161 published MTBC genomes including: 122 L2 strains (among these, 112 strains belonging to the Beijing sub-lineage); 37 L4 strains; and 2 L3 strains (6). We found that the missense mutation A809C (Glu270Ala) in *mamA* was present in all of the L2 strains, so it could be used in L2 lineage genotyping. This SNP has only been previously reported for typing within the L2.2 Beijing sub-lineage (43). Bioinformatic analysis also revealed that the missense mutation C917T (Pro306Leu) in *hsdM* was discovered in 35 out of the 37 L4 strains while, incidentally, the deletion within *mamB*, due to various forms of RvD1 in different strains, was present in the modern TB strains of L2, L3 and L4.

The evolution of the three MTBC DNA MTases

In order to explore the evolution of the three MTBC DNA MTases, we chose 1493 microbial species with complete genomes and built a phylogenetic tree for each MTase based on the protein sequences of each MTase. From Supplementary Figure S9A, we found that only 36 species contained MamA (Type II MTase), including all of the MTBC strains (four MTBC species), more than half of the non-tuberculosis mycobacteria (NTM) species (12 of 22 species), and *Mycobacterium leprae*, indicating that MamA might be indispensable to MTBC. Previous research reported that MamA could influence gene expression in *M. tuberculosis* (8). With regard to MamB MTase, there was a total of 60 species (including 4 MTBC species) which contained this (Supplementary Figure S9B). However, it was a surprise that MamB was not discovered in all of the 22 NTM species as these are considered to be the closest ancestors to MTBC (49). Since the MamB was discovered in many phyla in addition to *Actinobacteria*, such as *Proteobacteria*, *Firmicutes* and *Bacteroidetes*, we ascertained that the ancestor of NTM might have lost the MamB at an ancient evolutionary node (49). In comparison with MamA and MamB, HsdM MTase showed a much wider distribution across 302 species including 4 MTBC species, 6 NTM species and 292 other species (Supplementary Figure S9C). This indicates that HsdM MTase may be even more conserved than Type II MTases, because it may be evolutionarily constrained by the necessity of interacting with R and S subunits. Furthermore, we built a phylogenetic tree of the HsdS2 (S for specificity) recognition subunit as described previously (Supplementary Figure S10). Incidentally, in addition to the 12 MTBC strains that we sequenced, we also found the 3 MTase genes in all of the 36 published complete MTBC genomes, thereby suggesting that all the MTBC strains should contain the three MTase genes. It is interesting that the three DNA MTases are strongly conserved over MTBC strains, which needs further research.

DISCUSSION

Within this research, we have characterized the methylomes of 12 MTBC strains belonging to different MTBC lineages at single-base resolution using SMRT sequencing technology. Three ^{m6}A sequence motifs and their corresponding MTase genes were identified within MTBC strains. The evolution of the three MTases was also analyzed and discussed. Furthermore, through analyzing the 'the methylated-motif-site ratio' and 'the methylated-read ratio', we determined the precise methylome of the MTBC strains.

The precision methylome revealed the existence of unmethylated motif sites among the MTBC strains which had active MTases. To explore the formation mechanism of unmethylated motif sites, we studied the unmethylated sites that were found most frequently within the tested MTBC strains (Table 3). For both the G**A**TN$_4$R**T**AC and C**T**CC**A**G sites, the 10 most frequent unmethylated sites within gene regions appeared in at least 2 strains. Of particular interest, we found that three of the unmethylated sites occurred in all of the tested strains (Table 6), including two G**A**TN$_4$R**T**AC sites in pyrroline-5-carboxylate dehydrogenase gene (*rocA*) and cobalt-precorrin-6x reductase gene, respectively, and one C**T**CC**A**G site in the gene encoding for transmembrane transporter MmpL4. Supplementary Table S7 also revealed one unmethylated G**A**TN$_4$R**T**AC site in an intergenic region occurring in all of the tested strains. The appearance of these unmethylated sites across the tested strains increased the reliability of unmethylated site identification.

In order to investigate further why some sites always remained unmethylated, we analyzed the bind-

Table 6. The summary of top 10 frequent genes with unmethylated sites shared in 12 MTBC strains

Motif	Synonym	E/N	Gene annotation	MTBC Strain											
				F1	F28	26	30	12	25	2242	2279	22115	37004	22103	26105
GATN$_4$RTAC	Rv1187	Essential	pyrroline-5-carboxylate dehydrogenase RocA	NA	NA	-/-	-/-	-/+	-/-	-/-	-/-	NA	NA	-/-	NA
	Rv2070c	Non-essential	precorrin-6A reductase	NA	NA	-/-	-/-	-/-	-/-	-/+	-/-	NA	NA	-/-	NA
	Rv1753c	Essential	PPE family protein PPE24	NA	NA	-/-	-/-	+/+	-/-	-/-	-/-	NA	NA	-/-	NA
	Rv0112	Essential	GDP-mannose 4,6-dehydratase	NA	NA	-/-	-/-	-/+	-/-	-/-	+/+	NA	NA	-/+	NA
	Rv2963	Non-essential	integral membrane protein	NA	NA	-/-	-/-	+/+	-/-	-/-	-/-	NA	NA	-/-	NA
	Rv3341	Essential	homoserine O-acetyltransferase	NA	NA	-/-	-/-	+/+	-/-	-/-	-/+	NA	NA	-/-	NA
	Rv0405	Non-essential	membrane bound polyketide synthase	NA	NA	+/+	-/-	-/-	-/-	-/-	-/+	NA	NA	-/-	NA
	Rv1461	Essential	Fe-S cluster assembly protein SufB	NA	NA	-/-	-/-	-/+	-/-	+/+	-/+	NA	NA	-/-	NA
	Rv2130c	Essential	D-glucopyranoside ligase	NA	NA	-/-	-/-	+/+	-/-	+/+	+/+	NA	NA	+/+	NA
	Rv2984	No-data	polyphosphate kinase	NA	NA	-/-	-/-	+/+	-/-	+/+	+/+	NA	NA	+/+	NA
CTCCAG	Rv0450c	Essential	transmembrane transport protein MmpL4	-/-	-/-	-/-	-/-	-/-	-/-	NA	NA	-/-	-/-	-/-	-/-
	Rv1562c	No-data	malto-oligosyltrehalose trehalohydrolase	-/-	-/-	+/+	-/+	+/+	-/-	NA	NA	-/+	-/-	-/-	-/-
	Rv1461	Essential	Fe-S cluster assembly protein SufB	-/-	-/-	+/+	-/+	+/+	-/-	NA	NA	-/-	-/-	+/+	-/-
	Rv2501c	No-data	acetyl-/propionyl-CoA carboxylase subunit alpha	+/+	+/+	-/+	-/+	+/+	-/+	NA	NA	-/-	-/-	-/-	-/+
	Rv1917c	Non-essential	PPE family protein, PPE34	-/+	-/+	+/+	+/+	+/+	+/+	NA	NA	-/-	-/-	-/-	-/-
	Rv3282	Essential	Maf-like protein	+/+	-/+	+/+	+/+	+/+	+/+	NA	NA	-/-	-/-	-/+	-/+
	Rv0400c	Essential	acyl-CoA dehydrogenase FadE7	+/+	-/-	+/+	+/+	-/+	+/+	NA	NA	-/-	-/+	+/+	-/-
	Rv1664	Non-essential	polyketide synthase	+/+	+/+	-/+	+/+	+/+	+/+	NA	NA	-/-	-/+	-/+	+/+
	Rv2174	Essential	alpha-(1->6)-mannopyranosyltransferase A	+/+	+/+	+/+	+/+	+/+	+/+	NA	NA	-/-	-/-	-/+	-/+
	Rv1552	No-data	fumarate reductase flavoprotein subunit	+/+	+/+	+/+	+/+	+/+	+/+	NA	NA	-/-	-/+	+/+	+/+

E/N: essential / nonessential genes. NA: Not applicable. '-/-' denotes the gene with unmethylated motifs on both strands. '-/+' indicates gene with hemimethylated motif. '+/+' represents the gene with methylated motifs on both strands.

ing regions of 119 transcription factors (TF) from the TB Database (http://genome.tbdb.org/annotation/genome/tbdb/RegulatoryNetwork.html). We found that most of the unmethylated sites overlapped with TF binding regions (Supplementary Tables S7 and S8), and indeed some of them were overlapped with multiple TF binding regions (Supplementary Figure S11). From this finding we presumed that these sites were probably bound with some TFs, and thereby protected themselves from methylation (Supplementary Figure S12). Several publications concerning SMRT sequencing have also stated that some proteins, such as the Fur regulon in *Caulobacter crescentus* (50), might bind to certain motif sites in order to prevent methylation. This phenomenon is also commonly found in *E.coli*, in which some regulators (such as OxyR, SeqA) compete with Dam for occupation of a certain motif site (12). Alternatively, it is also possible that these unmethylated sites are protected from methylation by a higher-order chromatin structure (17).

The precision methylome of the MTBC strains also enables us to deeply investigate the activity of MTase at a whole-genome scale. Here, we discuss that relationship in more detail (Figures 2 and 3, Tables 2 and 4). For the CACGCAG motifs, the MTBC strains with active MamB displayed a 100% methylated-motif-site ratio, with most of them also displaying a high-level methylated-read ratio (90–100%), indicating that MamB was capable of high methylation activity within the MTBC strains. Of the CTCCAG motifs, all of the MTBC strains, except Mtb 37004, also displayed high methylation activities, shown by a high methylated-motif-site ratio (more than 99.5%) and a high-level methylated-read ratio (90–100%) in most motif sites. With respect to Mtb 37004, its methylated-motif-site ratio (∼97%) was a little lower than that of the other strains, but its average methylated-read ratio (82.06%) was noticeably lower than that of the other strains (95.40%) (Table 4). We presumed that a point mutation (G527A) in the catalytic domain of MamA (51) in Mtb 37004 might be the cause of this low methylation activity (Table 5B). In this instance, Mtb 37004 was revealed to be a good example of how to precisely analyze the activity of MTase by measuring the 'the methylated-motif-site ratio' and 'the methylated-read ratio'. For the GATN$_4$RTAC motifs, three ancient strains (*M. africanum* 25; *M. bovis* BCG 26; and *M. bovis* 30) exhibited lower methylated-motif-site ratios (82.43–90.41%) and an average methylated-read ratio (87.22%) by comparison with those of the other modern strains (94.78%). This might be due to the variety of HsdM MTase methylation activity within the different MTBC lineages.

In summary, this study showed the enormous potential of the PacBio SMRT platform for characterizing the 'precision methylome'. Not only could it detect the modified bases at single-base resolution, but it also could predict the modified sequence motif throughout the whole genome. Moreover, SMRT sequencing revealed that not all modified motif sites were detected as modified, and that there were a few unmodified motif sites. We determined 'the methylated-motif-site ratio' by calculating the percentage of modified motif sites within a total number of motif sites. SMRT sequencing was capable of even more. It could detect the modification status of each sequence read. Our findings demonstrated that, as seen for one modified motif site, not all reads covering the site were detected as modified. We obtained 'the methylated-read ratio' by calculating the percentage of modified reads within the total number of reads that had been mapped to the site. Overall, the characterization of the precision methylome significantly enhances our understanding of the function of DNA MTase.

To date, although the PacBio SMRT Sequencing technology has been used for the detection of many modified nucleotides throughout the whole genome, such as ^{m6}A, ^{m4}C, ^{m5}C, 5hmC and some damaged bases, its precision analysis for modification is still limited to ^{m6}A (17). This situation warrants further study in order to perfect the precise characterization of more modifications in the future.

ACCESSION NUMBERS

The SRA accession number for the sequencing data reported in this paper is SRP064893. The GenBank accession numbers are CP010329–CP010340.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

'100-Talent Program' of Chinese Academy of Sciences [Y3CAS81554]; Key Research Program of the Chinese Academy of Sciences [KJZD-EW-L02]; Infectious Diseases Special Project, Minister of Health of China [2013ZX10004605 and 2016ZX10003001-12]; Collaborative Innovation Center of Infectious diseases [PXM2015_014226_000058]. Funding for open access charge: '100-Talent Program' of Chinese Academy of Sciences [Y3CAS81554]; Key Research Program of the Chinese Academy of Sciences [KJZD-EW-L02]; Infectious Diseases Special Project, Minister of Health of China [2013ZX10004605 and 2016ZX10003001-12]; Collaborative Innovation Center of Infectious diseases [PXM2015_014226_000058].
Conflict of interest statement. None declared.

REFERENCES

1. World Health Organization. (2014) Global Tuberculosis Report 2014.
2. Rodriguez-Campos,S., Smith,N.H., Boniotti,M.B. and Aranaz,A. (2014) Overview and phylogeny of *Mycobacterium tuberculosis* complex organisms: Implications for diagnostics and legislation of bovine tuberculosis. *Res. Vet. Sci.*, **97**(Suppl), S5–S19.
3. Firdessa,R., Berg,S., Hailu,E., Schelling,E., Gumi,B., Erenso,G., Gadisa,E., Kiros,T., Habtamu,M., Hussein,J. *et al.* (2013) Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. *Emerg. Infect. Dis.*, **19**, 460–463.
4. Brosch,R., Gordon,S.V., Marmiesse,M., Brodin,P., Buchrieser,C., Eiglmeier,K., Garnier,T., Gutierrez,C., Hewinson,G., Kremer,K. *et al.* (2002) A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 3684–3689.
5. Ford,C.B., Shah,R.R., Maeda,M.K., Gagneux,S., Murray,M.B., Cohen,T., Johnston,J.C., Gardy,J., Lipsitch,M. and Fortune,S.M. (2013) *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat. Genet.*, **45**, 784–790.
6. Zhang,H., Li,D., Zhao,L., Fleming,J., Lin,N., Wang,T., Liu,Z., Li,C., Galwey,N., Deng,J. *et al.* (2013) Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat. Genet.*, **45**, 1255–1260.
7. Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E. 3rd *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.
8. Shell,S.S., Prestwich,E.G., Baek,S.H., Shah,R.R., Sassetti,C.M., Dedon,P.C. and Fortune,S.M. (2013) DNA methylation impacts gene expression and ensures hypoxic survival of *Mycobacterium tuberculosis*. *PLoS Pathog.*, **9**, e1003419.
9. Aird,D., Ross,M.G., Chen,W.S., Danielsson,M., Fennell,T., Russ,C., Jaffe,D.B., Nusbaum,C. and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.
10. Srivastava,R., Gopinathan,K.P. and Ramakrishnan,T. (1981) Deoxyribonucleic acid methylation in mycobacteria. *J. Bacteriol.*, **148**, 716–719.
11. Lederberg,S. (1966) Genetics of host-controlled restriction and modification of deoxyribonucleic acid in *Escherichia coli*. *J. Bacteriol.*, **91**, 1029–1036.
12. Wion,D. and Casadesus,J. (2006) N6-methyl-adenine: an epigenetic signal for DNA-protein interactions. *Nat. Rev. Microbiol.*, **4**, 183–192.
13. Casadesus,J. and Low,D. (2006) Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.*, **70**, 830–856.
14. Flusberg,B.A., Webster,D.R., Lee,J.H., Travers,K.J., Olivares,E.C., Clark,T.A., Korlach,J. and Turner,S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, **7**, 461–465.
15. Feng,Z., Li,J., Zhang,J.R. and Zhang,X. (2014) qDNAmod: a statistical model-based tool to reveal intercellular heterogeneity of DNA modification from SMRT sequencing data. *Nucleic Acids Res.*, **42**, 13488–13499.
16. Beaulaurier,J., Zhang,X.S., Zhu,S., Sebra,R., Rosenbluh,C., Deikus,G., Shen,N., Munera,D., Waldor,M.K., Chess,A. *et al.* (2015) Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. *Nat. Commun.*, **6**, 7438.
17. Fang,G., Munera,D., Friedman,D.I., Mandlik,A., Chao,M.C., Banerjee,O., Feng,Z., Losic,B., Mahajan,M.C., Jabado,O.J. *et al.* (2012) Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.*, **30**, 1232–1239.
18. Kozdon,J.B., Melfi,M.D., Luong,K., Clark,T.A., Boitano,M., Wang,S., Zhou,B., Gonzalez,D., Collier,J., Turner,S.W. *et al.* (2013) Global methylation state at base-pair resolution of the *Caulobacter* genome throughout the cell cycle. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, E4658–E4667.
19. Krebes,J., Morgan,R.D., Bunk,B., Sproer,C., Luong,K., Parusel,R., Anton,B.P., Konig,C., Josenhans,C., Overmann,J. *et al.* (2014) The complex methylome of the human gastric pathogen *Helicobacter pylori*. *Nucleic Acids Res.*, **42**, 2415–2432.
20. Lluch-Senar,M., Luong,K., Llorens-Rico,V., Delgado,J., Fang,G., Spittle,K., Clark,T.A., Schadt,E., Turner,S.W., Korlach,J. *et al.* (2013) Comprehensive methylome characterization of *Mycoplasma genitalium* and *Mycoplasma pneumoniae* at single-base resolution. *PLoS Genet.*, **9**, e1003191.
21. Fabre,M., Koeck,J.L., Le Fleche,P., Simon,F., Herve,V., Vergnaud,G. and Pourcel,C. (2004) High genetic diversity revealed by variable-number tandem repeat genotyping and analysis of hsp65 gene polymorphism in a large collection of "*Mycobacterium canettii*" strains indicates that the *M. tuberculosis* complex is a recently emerged clone of "*M. canettii*". *J. Clin. Microbiol.*, **42**, 3248–3255.
22. Constant,P., Perez,E., Malaga,W., Laneelle,M.A., Saurel,O., Daffe,M. and Guilhot,C. (2002) Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in the *Mycobacterium tuberculosis* complex. Evidence that all strains synthesize glycosylated p-hydroxybenzoic methyl esters and that strains devoid of phenolglycolipids harbor a frameshift mutation in the pks15/1 gene. *J. Biol. Chem.*, **277**, 38148–38158.
23. English,A.C., Richards,S., Han,Y., Wang,M., Vee,V., Qu,J., Qin,X., Muzny,D.M., Reid,J.G., Worley,K.C. *et al.* (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One*, **7**, e47768.
24. Lagesen,K., Hallin,P., Rodland,E.A., Staerfeldt,H.H., Rognes,T. and Ussery,D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.
25. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
26. Hyatt,D., Chen,G.L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
27. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.
28. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2015) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **43**, D298–D299.
29. Darling,A.E., Mau,B. and Perna,N.T. (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, **5**, e11147.
30. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.

31. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
32. Tarailo-Graovac,M. and Chen,N. (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. In: Andreas,DB (ed). *Curr. Protoc. Bioinformatics*. John Wiley & Sons, Inc., pp. 4.10.1–4.10.14.
33. Tamura,K., Stecher,G., Peterson,D., Filipski,A. and Kumar,S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.*, **30**, 2725–2729.
34. Kahramanoglou,C., Prieto,A.I., Khedkar,S., Haase,B., Gupta,A., Benes,V., Fraser,G.M., Luscombe,N.M. and Seshasayee,A.S. (2012) Genomics of DNA cytosine methylation in Escherichia coli reveals its role in stationary phase transcription. *Nat. Commun.*, **3**, 886.
35. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
36. Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.
37. Clark,T.A., Murray,I.A., Morgan,R.D., Kislyuk,A.O., Spittle,K.E., Boitano,M., Fomenkov,A., Roberts,R.J. and Korlach,J. (2012) Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res.*, **40**, e29.
38. Anton,B.P., Mongodin,E.F., Agrawal,S., Fomenkov,A., Byrd,D.R., Roberts,R.J. and Raleigh,E.A. (2015) Complete genome sequence of ER2796, a DNA methyltransferase-deficient strain of Escherichia coli K-12. *PLoS One*, **10**, e0127446.
39. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
40. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
41. Letunic,I. and Bork,P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.*, **39**, W475–W478.
42. Rogozin,I.B., Makarova,K.S., Natale,D.A., Spiridonov,A.N., Tatusov,R.L., Wolf,Y.I., Yin,J. and Koonin,E.V. (2002) Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res.*, **30**, 4264–4271.
43. Coll,F., McNerney,R., Guerra-Assuncao,J.A., Glynn,J.R., Perdigao,J., Viveiros,M., Portugal,I., Pain,A., Martin,N. and Clark,T.G. (2014) A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat. Commun.*, **5**, 4812.
44. Dean,F.B., Nelson,J.R., Giesler,T.L. and Lasken,R.S. (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.*, **11**, 1095–1099.
45. Piccolo,F.M. and Fisher,A.G. (2014) Getting rid of DNA methylation. *Trends Cell Biol.*, **24**, 136–143.
46. Loenen,W.A., Dryden,D.T., Raleigh,E.A. and Wilson,G.G. (2014) Type I restriction enzymes and their relatives. *Nucleic Acids Res.*, **42**, 20–44.
47. Murray,I.A., Clark,T.A., Morgan,R.D., Boitano,M., Anton,B.P., Luong,K., Fomenkov,A., Turner,S.W., Korlach,J. and Roberts,R.J. (2012) The methylomes of six bacteria. *Nucleic Acids Res.*, **40**, 11450–11462.
48. Seshasayee,A.S., Singh,P. and Krishna,S. (2012) Context-dependent conservation of DNA methyltransferases in bacteria. *Nucleic Acids Res.*, **40**, 7066–7073.
49. Rollat-Farnier,P.A., Santos-Garcia,D., Rao,Q., Sagot,M.F., Silva,F.J., Henri,H., Zchori-Fein,E., Latorre,A., Moya,A., Barbe,V. *et al.* (2015) Two host clades, two bacterial arsenals: evolution through gene losses in facultative endosymbionts. *Genome Biol. Evol.*, **7**, 839–855.
50. Gonzalez,D., Kozdon,J.B., McAdams,H.H., Shapiro,L. and Collier,J. (2014) The functions of DNA methylation by CcrM in *Caulobacter crescentus*: a global approach. *Nucleic Acids Res.*, **42**, 3720–3735.
51. Camus,J.C., Pryor,M.J., Médigue,C. and Cole,S.T. (2002) Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology*, **148**, 2967–7293.