ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Precision Methylome and *In Vivo* Methylation Kinetics Characterization of *Klebsiella pneumoniae*

Jing Fu [1,2,3,#], Ju Zhang [1,#], Li Yang [1,3,#], Nan Ding [1,#], Liya Yue [1,#], Xiangli Zhang [1,3], Dandan Lu [1,3], Xinmiao Jia [1,4], Cuidan Li [1], Chongye Guo [1], Zhe Yin [5], Xiaoyuan Jiang [5], Yongliang Zhao [3,6], Fei Chen [1,3,*], Dongsheng Zhou [5,*]

[1] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China*
[2] *Department of Oncology, Henan Provincial People's Hospital, People's Hospital of Zhengzhou University, People's Hospital of Henan University, Zhengzhou 450001, China*
[3] *University of Chinese Academy of Sciences, Beijing 100049, China*
[4] *Department of Medical Research Center, Peking Union Medical College Hospital, Peking Union Medical College & Chinese Academy of Medical Sciences, Beijing 100730, China*
[5] *State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing 100071, China*
[6] *CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China*

Received 10 February 2021; revised 19 March 2021; accepted 11 June 2021
Available 29 June 2021

Handled by Ting Wang

KEYWORDS

DNA methylation;
Epigenomics;
6mA;
5mC;
Epigenetic regulation

Abstract *Klebsiella pneumoniae* (*K. pneumoniae*) is an important pathogen that can cause severe hospital- and community-acquired infections. To systematically investigate its methylation features, we determined the whole-genome sequences of 14 *K. pneumoniae* strains covering varying serotypes, multilocus sequence types, clonal groups, viscosity/virulence, and drug resistance. Their methylomes were further characterized using Pacific Biosciences single-molecule real-time and bisulfite technologies. We identified 15 methylation motifs [13 *N*6-methyladenine (**6mA**) and two 5-methylcytosine (**5mC**) motifs], among which eight were novel. Their corresponding DNA methyltransferases were also validated. Additionally, we analyzed the genomic distribution of GATC and CCWGG methylation motifs shared by all strains, and identified differential distribution patterns of some hemi-/un-methylated GATC motifs, which tend to be located within intergenic

[*] Corresponding authors.
E-mail: chenfei@big.ac.cn (Chen F), dongshengzhou1977@gmail.com (Zhou D).
[#] Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2021.04.002

regions (IGRs). Specifically, we characterized the *in vivo* methylation kinetics at single-base resolution on a genome-wide scale by simulating the dynamic processes of replication-mediated passive demethylation and MTase-catalyzed re-methylation. The slow methylation of the G**A**TC motifs in the replication origin (*oriC*) regions and IGRs implicates the epigenetic regulation of replication initiation and transcription. Our findings illustrate the first comprehensive dynamic methylome map of *K. pneumoniae* at single-base resolution, and provide a useful reference to better understand **epigenetic regulation** in this and other bacterial species.

Introduction

Klebsiella pneumoniae (*K. pneumoniae*), an important member of the Enterobacteriaceae, can cause severe hospital- and community-acquired infections such as pneumonia, genitourinary tract infection, and septicaemia. There are various typing methods for *K. pneumoniae* strains, including serotyping, multilocus sequence typing (MLST), and clonal group (CG) typing [1,2]. Studies indicate that the hypervirulence phenotype usually corresponds to K1/K2/K57 serotypes and CG23-ST23 [1–3], while the multidrug resistance (MDR) phenotype often corresponds to CG258-ST11/ST258 [4].

Studies on DNA methylation in *K. pneumoniae* strains using molecular biological techniques identified three DNA methyltransferases (MTases) and corresponding motifs, including two restriction-modification (R-M) systems (M.KpnI: GGT**A**CC; M.KpnBI: CAA**A**N$_6$RTCA, where N = A or C or G or T and R = G or A) and one orphan MTase (Dam: G**A**TC) [5–7]. In each motif, the methylated nucleotide is shown in bold, and the nucleotide pairing with the methylated nucleotide on the complementary strand is marked with an underline. Further research on Dam revealed the epigenetic mechanism involved in regulating mismatch repair, virulence, and pathogenicity of *K. pneumoniae* strains [8].

Recent rapid progress on high-throughput sequencing techniques, such as Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing for accurate detection of modified bases [mainly *N*6-methyladenine (6mA)] on a genome-wide scale, and bisulfite sequencing for efficient analysis of genome-wide 5-methylcytosine (5mC) [9,10], has greatly facilitated DNA methylome investigations in bacteria. It is well known that 6mA and 5mC are the two most important types of DNA methylation in prokaryotes [9]. Numerous bacterial methylomes have been precisely determined using the aforementioned two techniques, including *Escherichia coli*, *Mycoplasma genitalium*, *Bifidobacterium breve*, *Clostridium difficile*, *Campylobacter jejuni*, *Helicobacter pylori* [9–13], and *Mycobacterium tuberculosis* complexes (MTBC) reported by our group [14].

By precisely and comprehensively analyzing the bacterial methylome, a lot of valuable information has been obtained, including methylation motifs and their corresponding MTases, motif distributions in genomes, and related epigenetic regulatory mechanisms in bacteria [9,15,16]. Most identified MTases and their corresponding motifs belong to the R-M system, which primarily functions to disrupt (cleave) invading DNA and protect genomic DNA through methylation-mediated mechanisms [9–14]. Distinctively, some orphan MTases (without cognate restriction enzymes) and their corresponding motifs perform multiple epigenetic regulatory functions in bacteria [15–18]. Among them, the Dam/G**A**TC motif is the most well-known due to its presence in almost all Enterobacteriaceae bacteria, and its involvement in the epigenetic regulation of replication, transcription, and mismatch repair [15–17,19–21]. In particular, its regulatory role in replication initiation has been studied in *E. coli*. Its replication origin (*oriC*) region contains five DnaA boxes and 11 G**A**TC motif sites. Replication-mediated passive demethylation causes the hemi-methylated G**A**TC motifs adjacent to the DnaA boxes to be specifically recognised and bound by SeqA, leading to competition for the motif sites between Dam and SeqA [17]. As a result, re-methylation of the motifs is delayed, which in turn prevents the initiation cascade for chromosome replication induced by the DnaA protein [17,19]. Re-methylation of the upstream G**A**TC motifs of the third and fifth DnaA boxes are the rate-limiting steps for DNA replication initiation in *E. coli* strains [17]. Additionally, Dam also participates in the transcriptional regulation of downstream genes such as *opvAB* in *Salmonella enterica* [15].

Although several MTases and corresponding motifs have been revealed in *K. pneumoniae* strains, the whole methylome has not yet been reported. Herein, we obtained whole-genome sequences of 14 *K. pneumoniae* strains of various types, and characterized their methylomes using SMRT/bisulfite sequencing. A total of 15 methylation motifs were identified, including 13 6mA and two 5mC methylation motifs. Among them, eight motifs were novel, corresponding to eight novel MTases [*K. pneumoniae* adenine methyltransferases A–G (KamA–G) and *K. pneumoniae* cytosine methyltransferase A (KcmA)]. We further analyzed the distribution patterns of the G**A**TC and C**C**WGG (where W = A or T) methylation motifs shared by all *K. pneumoniae* strains. Importantly, by establishing a mathematical model to simulate the dynamic processes of passive demethylation and re-methylation for each motif in the exponential phase, we characterized the genome-wide *in vivo* methylation kinetics at single-base resolution. Motifs at different genomic locations displayed different re-methylation rates, and the G**A**TC motifs in the *oriC* regions and intergenic regions (IGRs) had slow re-methylation rates. Our findings indicate potential roles of epigenetic regulation in replication initiation and transcription in the *K. pneumoniae* genome, and provide important insight into *K. pneumoniae* epigenomics.

Results

General bioinformatic analysis of 14 *K. pneumoniae* strains

We first obtained the whole-genome sequences of 14 *K. pneumoniae* strains (including NTUH-K2044, 11492, 11420, 11454, 12208, 11311, 23, 11305, N201205880, 309074, 13190, 283747, 721005, and 11021) differing in terms of serotypes,

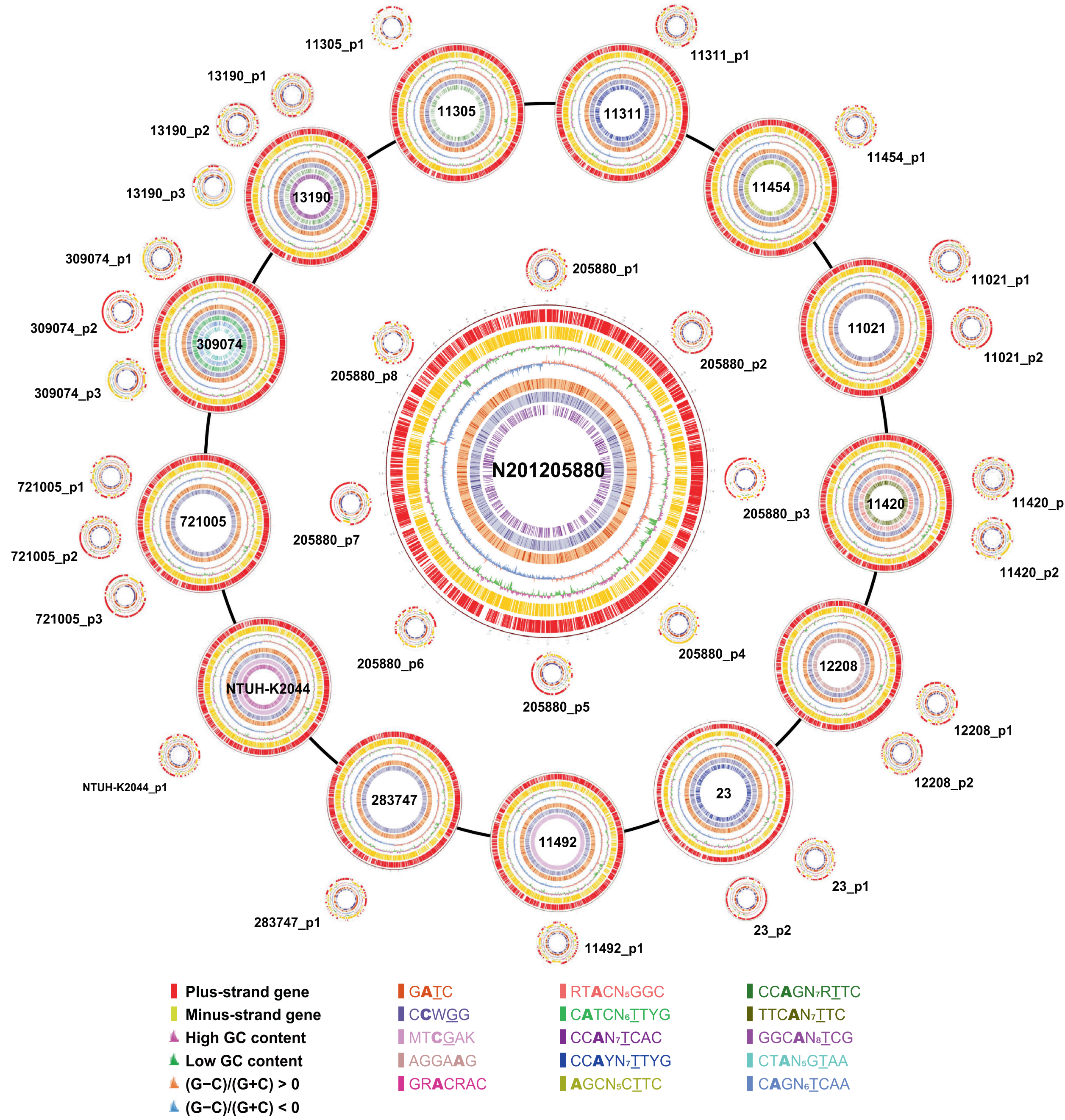

Figure 1 Circos plots displaying general genomic information for 14 *K. pneumoniae* strains

The circles are as follows (from outside to inside): 1) physical map scaled in megabases from base 1, the start of the putative replication origin; 2) coding sequences transcribed in a clockwise direction; 3) coding sequences transcribed in a counterclockwise direction; 4) GC content based on a 2-kb sliding window (red and green indicate a GC content higher or lower than average, respectively); 5) GC skew [(G−C)/(G + C)] based on a 2-kb sliding window (orange and green indicate GC skew above and below zero, respectively); 6) GATC motifs; and 7) CCWGG motifs. The other circles inside indicate the other 13 methylation motifs. In each motif, the methylated nucleotide is shown in bold, and the nucleotide pairing with the methylated nucleotide on the complementary strand is marked with an underline. Degenerate bases used in the motifs are as follows: R = G or A, Y = C or T, M = A or C, K = G or T, S = G or C, W = A or T, B = C or G or T, D = A or G or T, H = A or C or T, V = A or C or G, N = A or C or G or T. Mb, megabase.

MLSTs, CGs, viscosity/virulence, and drug resistance (Tables S1–S3) by SMRT sequencing followed by correction using Illumina sequencing (**Figure 1**). We then constructed a phylogenetic tree using 76 complete genomes of *K. pneumoniae* strains (14 from the current work and 62 obtained online) (Figure S1; Table S4). Our 14 *K. pneumoniae* strains covered many common CGs and MLSTs of *K. pneumoniae* strains in China (*e.g.*, CG23-ST23 and CG258-ST11) [22], indicating good representation and diversity of *K. pneumoniae* strains selected in our study.

Table 1 General genomic information for the 14 *K. pneumoniae* strains

Strain name	Chromosome/plasmid	Genomic size (Mb)	GC content (%)	No. of genes	Average gene size (bp)	Coding region (%)	No. of tRNAs	No. of rRNAs
NTUH-K2044	NUTH-K2044_chr	5.25	57.68	5050	917	90.09	86	25
	pNUTH-K2044-1	0.22	50.17	243	701	83.28	0	0
11492	11492_chr	5.25	57.45	5017	922	88.43	87	25
	p11492-1	0.19	50.37	202	724	82.72	0	0
11420	11420_chr	5.44	57.46	5264	912	88.35	88	25
	p11420-1	0.23	50.10	257	688	83.43	0	0
	p11420-2	0.08	50.56	101	605	84.45	0	0
11454	11454_chr	5.20	57.60	4990	917	88.35	86	25
	p11454-1	0.21	50.67	225	717	83.74	0	0
12208	12208_chr	5.21	57.59	5067	910	89.42	82	25
	p12208-1	0.33	47.27	353	709	83.79	0	0
	p12208-2	0.21	49.69	232	698	82.90	0	0
11311	11311_chr	5.27	57.56	5048	918	88.26	86	25
	p11311-1	0.20	49.93	209	733	82.74	0	0
23	23_chr	5.29	57.54	5120	910	88.31	85	25
	p23-1	0.20	49.93	212	730	82.76	0	0
	p23-2	0.11	49.46	106	831	87.78	2	0
11305	11305_chr	5.32	57.46	5186	901	88.37	88	25
	p11305-1	0.06	53.29	60	690	75.98	0	0
N201205880	N201205880_chr	5.21	57.50	5106	901	88.85	87	25
	pN201205880-1	0.23	51.08	240	758	82.03	0	0
	pN201205880-2	0.15	52.33	146	755	88.63	0	0
	pN201205880-3	0.11	54.85	134	661	83.30	0	0
	pN201205880-4	0.11	49.31	108	815	88.07	1	0
	pN201205880-5	0.08	50.29	88	733	84.69	0	0
	pN201205880-6	0.08	54.00	96	596	78.08	0	0
	pN201205880-7	0.07	52.85	83	657	86.33	0	0
	pN201205880-8	0.06	52.45	69	609	82.19	0	0
309074	309074_chr	5.31	57.46	5181	907	90.23	89	25
	p309074-1	0.19	53.44	197	735	83.60	0	0
	p309074-2	0.07	51.83	85	668	83.78	0	0
	p309074-3	0.05	49.03	65	667	85.67	0	0
13190	13190_chr	5.26	57.50	5172	897	89.34	88	25
	p13190-1	0.31	47.26	316	703	84.09	0	0
	p13190-2	0.08	53.86	85	649	79.42	0	0
	p13190-3	0.04	45.72	50	674	86.66	0	0
283747	283747_chr	5.50	57.37	5441	891	90.07	85	25
	p283747-1	0.15	54.40	192	574	78.82	0	0
721005	721005_chr	5.48	57.38	5468	886	89.57	85	25
	p721005-1	0.27	46.48	283	716	82.85	0	0
	p721005-2	0.16	53.92	199	627	82.14	0	0
	p721005-3	0.06	51.76	77	631	88.59	0	0
11021	11021_chr	5.45	57.46	5386	893	89.67	85	25
	p11021-1	0.24	51.69	271	682	82.43	0	0
	p11021-2	0.17	53.48	217	605	81.85	0	0

The bioinformatic analyses provided general genome information (**Table 1**, Figure 1), including genome size (5.20–5.50 Mb), GC content (57.37%–57.68%), predicted number of protein-coding genes (4990–5468), gene length (886–922 bp), and the percentage of coding regions (88.26%–90.23%) [23]. In addition, each *K. pneumoniae* strain contained 1–8 plasmids, with lower GC content (45.72%–54.85%), lower percentage of coding regions (75.98%–88.63%), and shorter average gene length (574–831 bp) (Table 1).

Additionally, the relatively conserved genomic sequences and structures among the 14 *K. pneumoniae* strains were also indicated. Average nucleotide identity (ANI) analysis revealed more than 99% identity among the 14 *K. pneumoniae* strains; no extensive translocations, duplications, or inversions were found in the *K. pneumoniae* genomes except for strains 11492 and 11454, each containing a large inverted fragment in the genome (Figure S2).

Seven known and eight novel methylation motifs and corresponding MTases in *K. pneumoniae* strains

A total of 13 6mA and two 5mC methylation motifs were identified in the 14 *K. pneumoniae* strains (**Tables 2–4**) by SMRT and bisulfite sequencing techniques (Tables S2 and S5), including seven known and eight novel methylation motifs (**Table 5**, Table S6). It is worth noting that GA$\underline{\text{T}}$C and CCW$\underline{\text{G}}$G motifs

Table 2 The three types of modification patterns of the GATC motif in the 14 *K. pneumoniae* strains

Strain name	Motif	No. of motifs	Methylated motif (%)	Hemi-methylated motif (%)	Un-methylated motif (%)
NTUH-K2044	GATC	30,727	30,151 (98.13%)	551 (1.79%)	25 (0.08%)
11492	GATC	30,316	30,271 (99.85%)	31 (0.10%)	14 (0.05%)
11420	GATC	31,847	26,835 (84.26%)	4669 (14.66%)	343 (1.08%)
11454	GATC	30,278	28,046 (92.63%)	2142 (7.07%)	90 (0.30%)
12208	GATC	31,512	30,695 (97.41%)	790 (2.51%)	27 (0.08%)
11311	GATC	30,623	29,520 (96.4%)	1076 (3.51%)	27 (0.09%)
23	GATC	31,275	26,523 (84.81%)	4423 (14.14%)	329 (1.05%)
11305	GATC	30,369	29,222 (96.22%)	1122 (3.69%)	25 (0.09%)
N201205880	GATC	33,171	28,007 (84.43%)	4852 (14.63%)	312 (0.94%)
309074	GATC	31,208	30,384 (97.36%)	782 (2.51%)	42 (0.13%)
13190	GATC	31,419	28,766 (91.56%)	2515 (8.00%)	138 (0.44%)
283747	GATC	31,335	30,156 (96.24%)	1126 (3.59%)	53 (0.17%)
721005	GATC	32,242	31,521 (97.76%)	687 (2.13%)	34 (0.11%)
11021	GATC	32,080	27,907 (86.99%)	3968 (12.37%)	205 (0.64%)

Note: The methylated adenine in the motif is shown in bold; the underlined letter represents the thymine pairing with the methylated adenine on the complementary strand. The number of motifs include the ones on the plus and minus strands of chromosomes and plasmids.

Table 3 Modification patterns of the 12 motifs with 6mA in the 14 *K. pneumoniae* strains

Strain name	Motif	No. of motifs	Methylated motif (%)	Hemi-methylated motif (%)	Un-methylated motif (%)
NTUH-K2044	GRACRAC*	2104	2060 (97.91%)	/	44 (2.09%)
11420	RTACN$_5$GGC*	1275	1156 (90.67%)	/	119 (9.33%)
	TTCAN$_7$TTC*	853	598 (70.11%)	231 (27.08%)	24 (2.81%)
11454	AGCN$_5$CTTC	991	930 (93.84%)	61 (6.16%)	/
12208	AGGAAG*	2864	2848 (99.30%)	/	16 (0.7%)
11311#	CCAYN$_7$TTYG*	667	608 (91.15%)	58 (8.70%)	1 (0.15%)
23#	CCAYN$_7$TTYG*	687	534 (77.73%)	145 (21.11%)	8 (1.16%)
11305^	CCAGN$_7$RTTC	342	325 (95.03%)	17 (4.97%)	/
N201205880	CCAN$_7$TCAC*	543	433 (79.75%)	102 (18.78%)	8 (1.47%)
309074	CTAN$_5$GTAA	165	164 (99.39%)	1 (0.61%)	/
	CAGN$_6$TCAA*	439	413 (94.08%)	26 (5.92%)	/
	CATCN$_6$TTYG	634	589 (92.90%)	45 (7.10%)	/
13190^	CCAGN$_7$RTTC	379	338 (89.19%)	40 (10.55%)	1 (0.26%)
	GGCAN$_8$TCG	1065	874 (82.06%)	184 (17.28%)	7 (0.66%)

Note: The methylated adenine in each motif is shown in bold; the underlined letter represents the thymine pairing with the methylated adenine on the complementary strand. Degenerate bases used in our recognition sequences are as follows: R = G or A, Y = C or T, M = A or C, K = G or T, S = G or C, W = A or T, B = not A (C or G or T), D = not C (A or G or T), H = not G (A or C or T), V = not T (A or C or G), N = A or C or G or T. The number of motifs include the ones on the plus and minus strands of chromosomes and plasimds. *, newly reported motifs. #, two strains (11311 and 23) possess motif CCAYN$_7$TTYG. ^, two strains (11305 and 13190) possess motif CCAGN$_7$RTTC.

were shared by all strains. The other motifs were shared by at most two *K. pneumoniae* strains (Tables 2–4). Further analyses indicated the relationships between the motifs and stain types (serotype, MLST, and CG). The MTCGAK (where M = A or C, K = G or T) motif was present in NTUH-K2044 and 11492 belonging to the K1 serotype and ST23-CG23, the most common types of hypervirulent *K. pneumoniae* strains. The CCAYN$_7$TTYG (where Y = C or T, N = A or C or G or T) motif was shared by two strains (11311 and 23) of hypervirulent serotype K57 and ST412. The CCAGN$_7$RTTC (where N = A or C or G or T, R = G or A) motif was present in strains 11305 and 13190 belonging to MDR CG147.

Modification analysis indicated that not all motif sites were fully methylated (methylated on both strands, Tables 2–4). A minority of motif sites ($<$ 30%) were detected as being hemi-methylated (methylated on one strand only) or un-methylated within the *K. pneumoniae* genomes. The only exception was the MTCGAK motif in the NTUH-K2044 and 11492 genomes, in which over half of the sites were hemi-/un-methylated (56.03%–56.77%). Further analysis indicated that the un-methylated MTCGAK motif tended to be preceded by a guanine (G) (Figure S3, Table S7).

To search for the respective MTases, we first predicted 22 MTase genes [24]. Among them, seven genes encode MTases which as well as their corresponding motifs had been verified in previous studies (Table 5). To determine the MTases responsible for the eight newly detected methylation motifs, we performed restriction digestion and SMRT/bisulfite sequencing using plasmids containing the predicted MTase genes in MTase-free *E. coli* ER2796. Crossover validation identified the corresponding eight MTases that could specifically recognize and methylate the respective eight novel motifs (Table 5, Tables S6 and S8; Figures S4–S6).

We further analyzed the distribution of 15 MTase genes in the genomes of the *K. pneumoniae* strains. Thirteen genes were located on chromosomes, and two others were located on

Table 4 Modification patterns of the two motifs with 5mC in the 14 *K. pneumoniae* strains

Strain name	Motif	No. of motifs	Methylated motif (%)	Hemi-methylated motif (%)	Un-methylated motif (%)
NTUH-K2044	CCWGG	19,530	16,117 (82.52%)	1756 (8.99%)	1657 (8.49%)
	MTCGAK*	5410	2339 (43.23%)	270 (4.99%)	2801 (51.78%)
11492	CCWGG	19,284	19,015 (98.61%)	86 (0.45%)	183 (0.94%)
	MTCGAK*	5386	2368 (43.97%)	258 (4.79%)	2760 (51.24%)
11420	CCWGG	20,175	18,828 (93.32%)	839 (4.16%)	508 (2.52%)
11454	CCWGG	19,314	14,983 (77.58%)	2367 (12.26%)	1964 (10.16%)
12208	CCWGG	20,108	19,874 (98.84%)	35(0.17%)	199 (0.99%)
11311	CCWGG	19,668	19,126 (97.24%)	299 (1.52%)	243 (1.24%)
23	CCWGG	19,941	18,668 (93.62%)	963 (4.83%)	310 (1.55%)
11305	CCWGG	19,460	18,964 (97.45%)	261 (1.34%)	235 (1.21%)
N201205880	CCWGG	21,115	20,867 (98.83%)	38 (0.18%)	210 (0.99%)
309074	CCWGG	19,966	17,464 (87.47%)	1460 (7.31%)	1042 (5.22%)
13190	CCWGG	20,060	19,149 (95.46%)	694 (3.46%)	217 (1.08%)
283747	CCWGG	20,156	19,141 (94.96%)	574 (2.85%)	441 (2.19%)
721005	CCWGG	20,832	20,470 (98.26%)	145 (0.70%)	217 (1.04%)
11021	CCWGG	20,736	18,352 (88.50%)	1358 (6.55%)	1026 (4.95%)

Note: The methylated cytosine in the motif is shown in bold; the underlined letter represents the guanine pairing with the methylated cytosine on the complementary strand. Degenerate bases used in our recognition sequences are listed as follows: R = G or A, Y = C or T, M = A or C, K = G or T, S = G or C, W = A or T, B = not A (C or G or T), D = not C (A or G or T), H = not G (A or C or T), V = not T (A or C or G), N = A or C or G or T. The number of motifs include the ones on the plus and minus strands of chromosomes and plasimds. *, newly reported motifs.

Table 5 The 15 methylation motifs and corresponding DNA MTases in the 14 *K. pneumoniae* strains

Motif	Modification	R-M/orphan	MTase	Comment
GATC	6mA	Orphan	Dam	Known
RTACN$_5$GGC	6mA	R-M system Type I	KamA	New
TTCAN$_7$TTC	6mA	R-M system Type I	KamB	New
GRACRAC	6mA	R-M system Type II	KamC	New
AGCN$_5$CTTC	6mA	R-M system Type I	M.KpnGH01II	Known
AGGAAG	6mA	R-M system Type II	KamD	New
CCAYN$_7$TTYG	6mA	R-M system Type I	KamE	New
CCAN$_7$TCAC	6mA	R-M system Type I	KamF	New
CCAGN$_7$RTTC	6mA	R-M system Type I	M.KpnAATI	Known
CTAN$_5$GTAA	6mA	R-M system Type I	M.Kpn35657I	Known
CAGN$_6$TCAA	6mA	R-M system Type I	KamG	New
CATCN$_6$TTYG	6mA	R-M system Type I	M.Kpn39795II	Known
GGCAN$_8$TCG	6mA	R-M system Type I	M.KpnAATIV	Known
CCWGG	5mC	Orphan	Dcm	Known
MTCGAK	5mC	R-M system Type II	KcmA	New

Note: The methylated nucleotide in each motif is shown in bold; the underlined letter represents the nucleotide pairing with the methylated nucleotide on the complementary strand. Degenerate bases used in the recognition sequences are as follows: R = G or A, Y = C or T, M = A or C, K = G or T, S = G or C, W = A or T, B = not A (C or G or T), D = not C (A or G or T), H = not G (A or C or T), V = not T (A or C or G), N = A or C or G or T. The MTase prediction was based on the sequence alignment with REBASE database (http://rebase.neb.com/rebase/rebase.html). The predicted MTases were further classified as Type I, Type II, or orphan MTases according to the annotation information. MTase, methyltransferase; R-M, restriction-modification.

plasmids (**Figure 2**). Additionally, within the 15 identified MTases, there were 10 Type I MTases, three Type II MTases, and two classical orphan MTases (Dam and Dcm). Here, *dam* and *dcm* genes were present in all *K. pneumoniae* strains, and responsible for the methylation of GATC and CCWGG motifs, respectively. Among the three newly identified Type II MTases, KamC and KamD were predicted to be Type IIG enzymes, for which endonuclease and methyltransferase activities are encoded by a single gene (Table 5, Table S6).

Nonrandom distributions of GATC and CCWGG motifs in *K. pneumoniae* genomes

Among the 15 methylation motifs, GATC and CCWGG had the most extensive distributions in all 14 *K. pneumoniae* strains, each containing ~ 30,000 GATC and ~ 20,000 CCWGG modified sites (Tables 2 and 4). The distributions of the two motifs in the genomes of *K. pneumoniae* strains were further analyzed, and differential/uneven distributions were observed (Figure 1). Both motifs exhibited some high-density and low-density regions in the genomes, where the genes were clustered into different Clusters of Orthologous Groups (COG) functional categories (Figure S7). Notably, the GATC motif showed the highest distribution density in the *oriC* region (~ 34 sites/kb) of the 14 *K. pneumoniae* genomes (average density: 5–6 sites/kb) (Figure S8). By contrast, the CCWGG motif did not display such enrichment in the *oriC* region.

We then compared the density distributions of the two motifs in the 14 *K. pneumoniae* genomes and the simulated genome with the same base composition (**Figure 3**, Figures S9 and

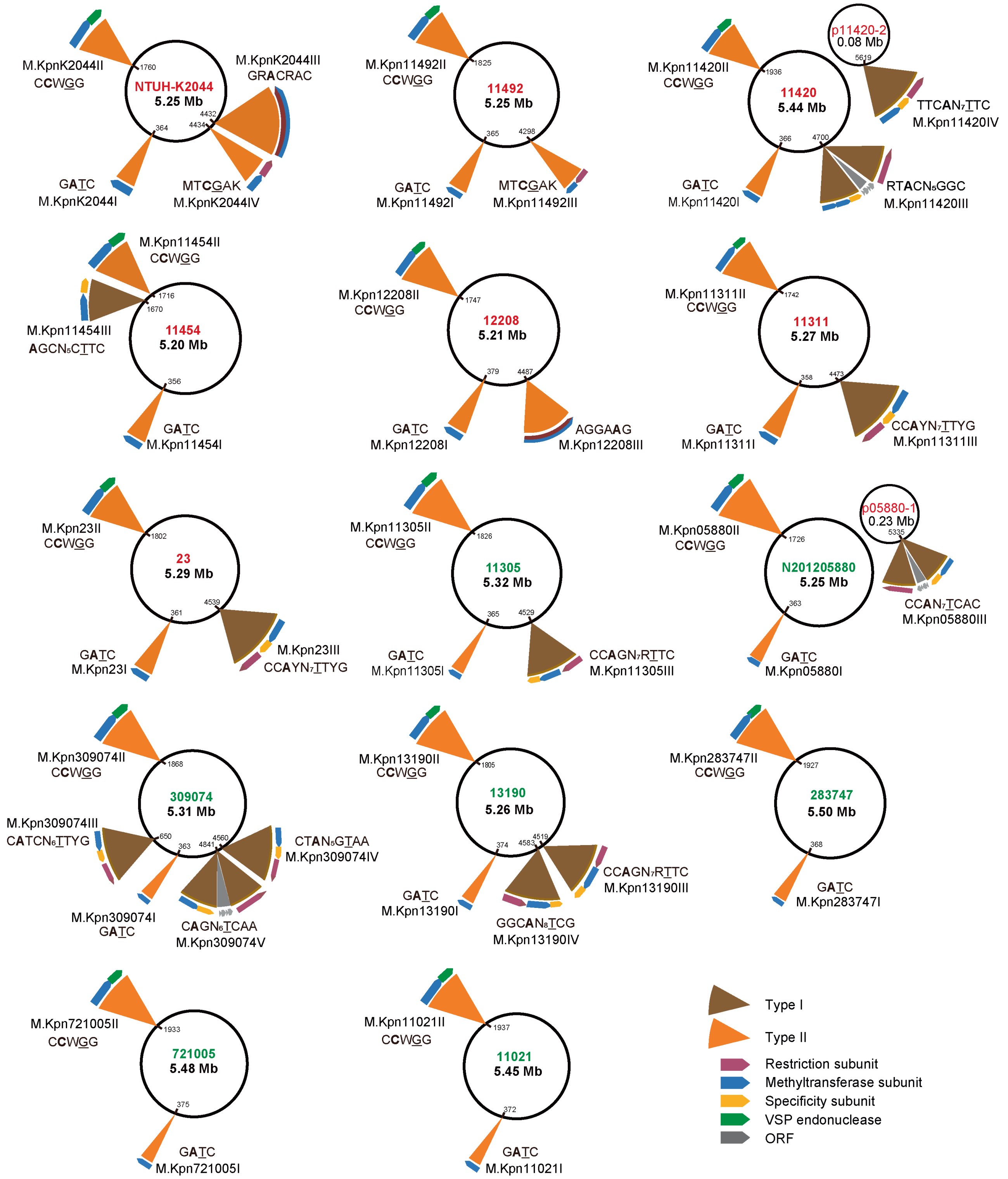

Figure 2 The 15 MTase genes and corresponding methylation motifs for the 14 *K. pneumoniae* strains
Hypervirulent and low-virulent strains are labeled in red and green, respectively. The large and small circles represent chromosome and plasmid genomes, respectively. ORF, open reading frame.

S10). The results revealed that their distribution densities in the *K. pneumoniae* genomes (1-kb consecutive window) were higher than those in the simulated genome, indicating a high-density/nonrandom distribution for these two motifs in the *K. pneumoniae* genome. To explore the underlying causes for their high-density distributions, we investigated the impact

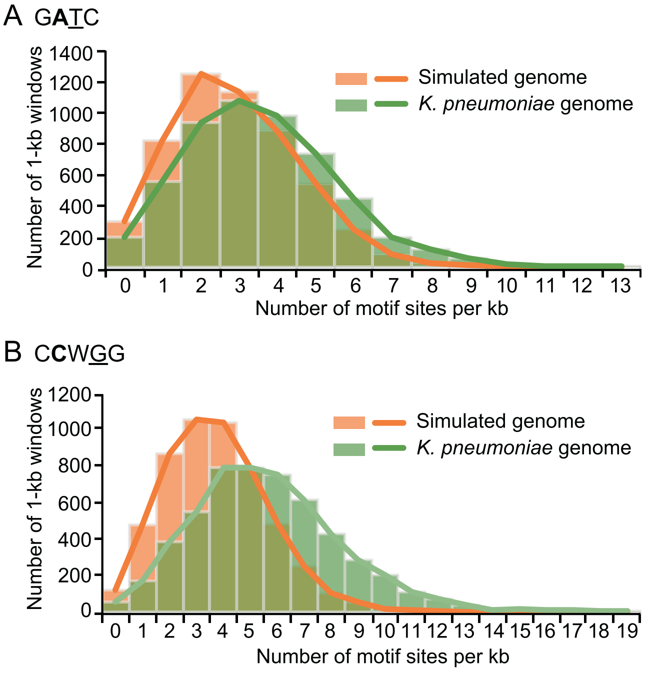

Figure 3 Density distribution of the GATC/CCWGG motifs in the *K. pneumoniae* and randomly generated genomes

Green histograms show the density distributions of GATC (**A**) and CCWGG (**B**) in the *K. pneumoniae* genome, which follow Poisson distributions with $\lambda = 5.64/3.62$. Orange histograms show the density distributions of GATC (A) and CCWGG (B) motifs in the randomly generated genome, which also follow Poisson distributions with $\lambda = 3.73/2.89$ (total number of GATC/CCWGG motifs × 1000 per genome size).

of selection pressure on these two motifs by calculating the ratio of nonsynonymous substitutions (Ka) to synonymous substitutions (Ks) [25] of the corresponding fragments in gene regions (GRs), and more than 90% of the two motifs were located in GRs (Figure S11). We observed that the amino acid (AA) codons with two motifs (two AA codons for the GATC motif, two or three AA codons for the CCWGG motif) were under strong negative/purifying selection with Ka/Ks of ∼ 0.09/0.09, compared with a ratio of 0.39/0.54 for the scramble (control) sequences in GRs.

Differential distribution patterns of methylated, hemi-methylated, and un-methylated GATC and CCWGG motifs in *K. pneumoniae* genomes

We identified three methylation patterns (methylated, hemi-methylated, and un-methylated) for the GATC and CCWGG motifs. Most GATC (84.26%–99.85%) and CCWGG (77.58%–98.84%) sites were found to be methylated (Tables 2 and 4), while only a small percentage of motif sites were hemi-methylated ($< 15\%$) or un-methylated ($< 10\%$) in the *K. pneumoniae* genomes. Further analysis demonstrated that the ratio of hemi-methylated GATC motifs (∼ 6.48%) was much higher than that of un-methylated GATC motifs (∼ 0.38%), while hemi-methylated (∼ 3.60%) and un-methylated (∼ 2.94%) CCWGG motifs accounted for similar proportions.

We then investigated the distribution ratios of methylated, hemi-methylated, and un-methylated GATC and CCWGG motifs in GRs and IGRs (Tables S9 and S10). The hemi-/un-methylated GATC motifs tended to be located in IGRs, since their ratios in IGRs were significantly higher than that of the methylated GATC motifs (7.78% for hemi-methylated, 41.79 for un-methylated, 5.14% for methylated; $P < 0.01$, **Figure 4**A, Table S9). Analysis of the 'fraction of methylated reads' (FRAC value) for the motifs in GRs/IGRs also supported the aforementioned finding that hemi-/un-methylated GATC motifs tended to be located in the IGRs (Figure 4B).

Analysis of the GATC motif density in the 5′ upstream region (USR), the coding sequence (CDS), and the 3′ downstream region (DSR) also supported the aforementioned conclusion: the methylated motifs displayed a higher density in GRs, while the hemi-/un-methylated motifs were more abundant in IGRs (both 5′ USR and 3′ DSR; Figure 4C). As for the CCWGG motifs, we did not observe higher abundant hemi-/un-methylated sites in IGRs than in GRs (Figure 4D).

We further explored the hemi-/un-methylated GATC sites shared in the 5′ USRs of 14 *K. pneumoniae* strains, and 13 hemi-/un-methylated GATC sites corresponding to 11 genes were observed (Table S11). We also detected 12 high-density hemi-/un-methylated GATC clusters (no less than three consecutive hemi-/un-methylated motifs in at least two strains) in the 5′ USRs (Table S12).

We then analyzed the sequence conservation of the DNA fragments (20 nt on both sides) containing hemi-/un-methylated or methylated motifs to determine the causes leading to their differential distributions in IGRs. The results showed that the fragments with hemi-/un-methylated GATC motifs (44 nt) exhibited higher conservation than those with methylated GATC motifs in IGRs (Figure S12).

Methylation kinetic analysis revels different re-methylation rates for GATC and CCWGG motifs during the growth cycle

Since there are no demethylases in bacteria, *in vivo* methylation kinetics characterisation is based on the dynamic equilibrium between replication-mediated passive demethylation and MTase-catalyzed re-methylation [18]. To explore the features of methylation kinetics of GATC and CCWGG motifs during the growth cycle, we first characterized the methylomes of the NTUH-K2044 and 11492 strains at the exponential phase (at 1 h), the transition-to-stationary phase (4 h), and the stationary phase (at 24 h) (**Figure 5**A; Tables S13 and S14). By comprehensively analyzing the genome-wide sequencing coverage and the fraction of methylated reads (methylated read ratio/FRAC value), GATC and CCWGG motifs were found to exhibit distinct kinetic features during the growth cycle (Figure 5B, Figure S13). Regarding the GATC motif, the methylated read ratios were more than 90% throughout the genomes in all three phases (Figure 5B), although the sequencing coverage varied between genomes during the growth cycle (Figure S13A). This indicates that the GATC motif may be re-methylated in a very short time after passive demethylation

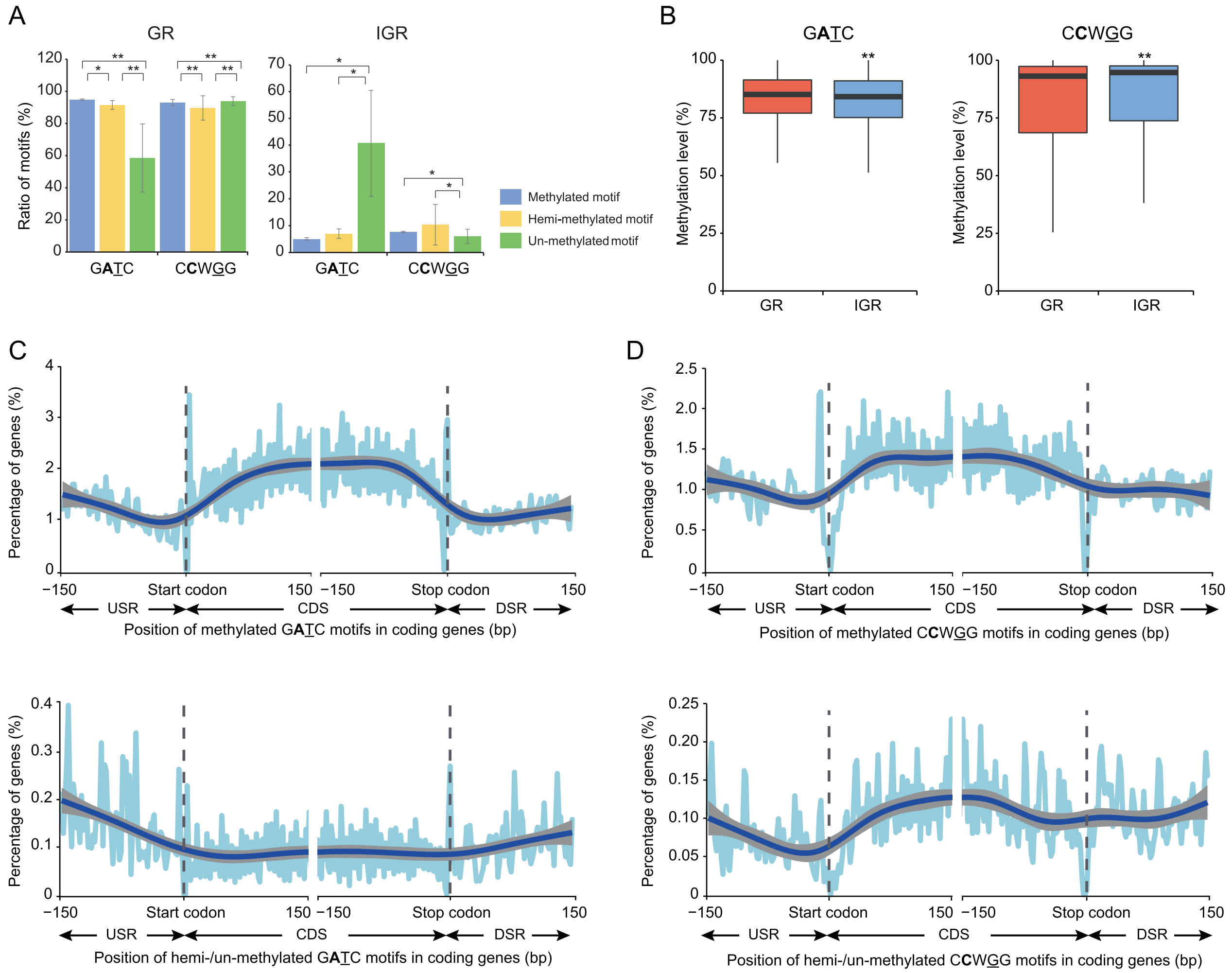

Figure 4 Distribution of GATC/CCWGG motifs with different methylation patterns in GRs and IGRs

A. Bar plots showing the ratios of GATC/CCWGG motifs with different methylation patterns in GRs and IGRs in the 14 *K. pneumoniae* strains. Blue, yellow, and green bars indicate the ratios of methylated, hemi-methylated, and un-methylated motifs in GRs and IGRs, respectively. **B.** Box plots showing the methylation levels of GATC/CCWGG motifs in GRs and IGRs (red and blue boxes, respectively). **C.** Frequency distribution of the methylated (top) and hemi-/un-methylated (bottom) GATC motifs in GRs and IGRs in the 14 *K. pneumoniae* strains. **D.** Frequency distribution of the methylated (top) and hemi-/un-methylated (bottom) CCWGG motifs in GRs and IGRs in the 14 *K. pneumoniae* strains. The dotted gray lines represent the positions of the start and stop codons. USR, 5′ upstream region; CDS, coding sequence; DSR, 3′ downstream region; GR, gene region; IGR, intergenic region. *, $P < 0.05$; **, $P < 0.01$.

caused by replication (*i.e.*, the re-methylation rate was almost identical to the passive demethylation rate).

Compared to the GATC motif, the CCWGG motif showed a much slower re-methylation. First, in the exponential phase (at 1 h), the methylated read ratio of CCWGG motifs in the *oriC* regions (55.66% ± 18.54%) was much lower than that in the replication termination (*Ter*) regions (82.63% ± 12.71%; Figure S13B). Secondly, the average methylated read ratio in the transition-to-stationary phase (54.50% ± 27.14%, at 4 h) was close to that of the exponential phase (60.50% ± 17.71%, at 1 h), but not close to that of the stationary phase (80.28% ± 19.05%, at 24 h; Figure S13B). These findings suggest a much slower re-methylation, compared to the replication-mediated passive demethylation rate for the CCWGG motif, which resulted in the aforementioned differences in methylated read ratios.

To quantify the re-methylation time per motif in the exponential phase, we first investigated the replication kinetics of the two aforementioned *K. pneumoniae* strains. Figure S13 shows that the genome sequencing coverage in the *oriC* regions was more than twice as much as that in the *Ter* regions (2.93- and 3.02-fold for the 11492 and NTUH-K2044 strains, respectively), indicating that the next round of replication was re-initiated before replication termination in the exponential phase. We then constructed a replication model (Figure 5C) by fitting genome coverage data, and obtained the ratio (t_D/t_R) of doubling time (t_D) to replication termination time (t_R) as described in the Materials and methods. The t_D in the exponential phase also reflects the re-initiation time. Interestingly, we obtained the same t_D/t_R values (0.59) for the two *K. pneumoniae* strains (Figure 5D), suggesting a similar regulatory mechanism for the replication cycle, as expected for

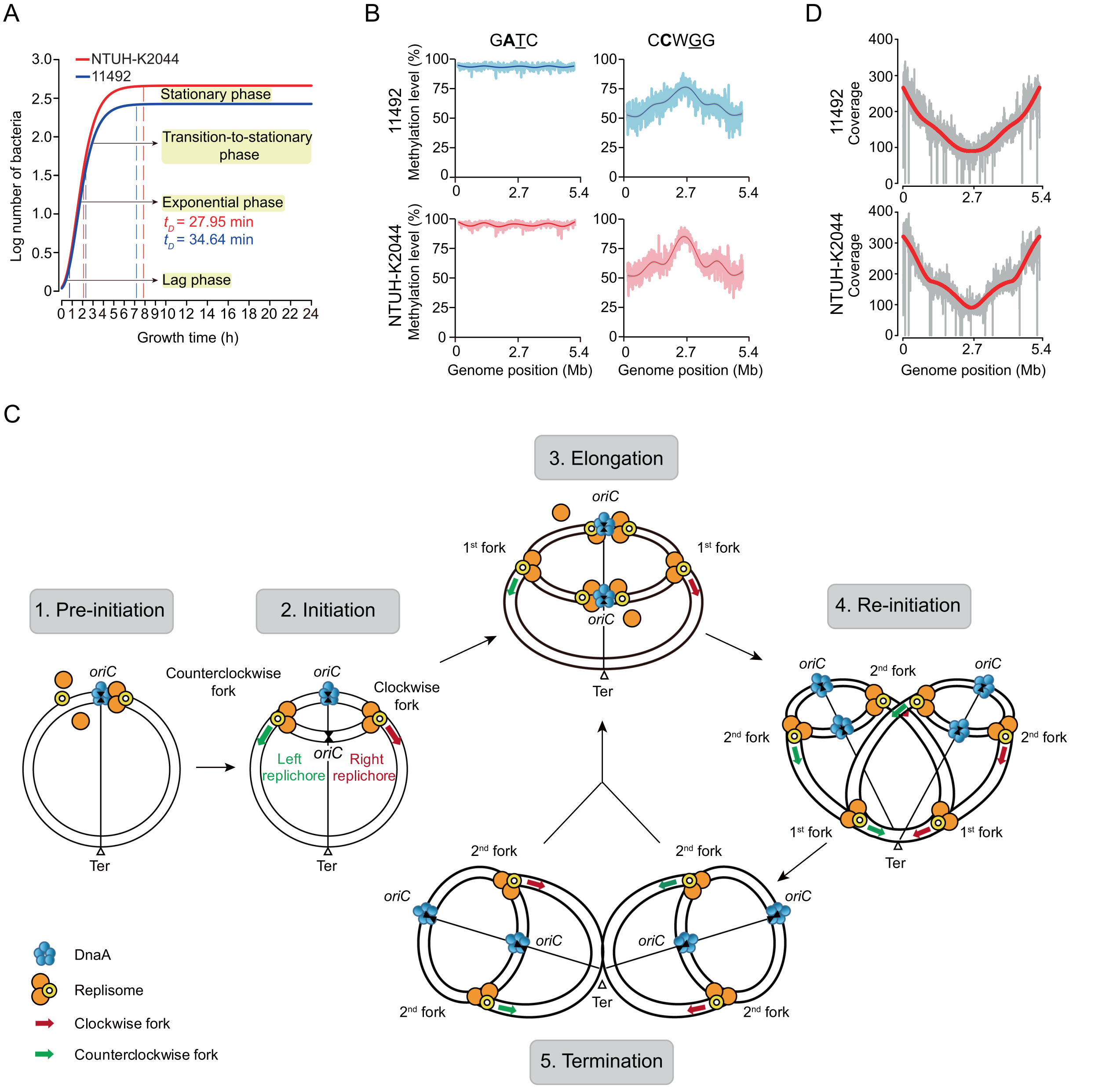

Figure 5 Dynamic methylation analysis of GATC and CCWGG motifs in NTUH-K2044 and 11492

A. Growth curves of *K. pneumoniae* NTUH-K2044 (red) and 11492 (blue). The X-axis represents the growth time, including lag phase, exponential phase, transition-to-stationary phase, and stationary phase. The Y-axis represents the logarithms of normalized OD values. At 1, 4, and 24 h, aliquots of bacterial cultures were collected and sequenced using PacBio and Illumina platforms. The t_D values of the two strains are also labeled in the plot. **B.** Genome-wide methylation level *versus* genome position for the two motifs (GATC and CCWGG) at the exponential phase. The bold fitting lines approximate the average methylation levels across the genomes (5-kb window size). **C.** Schematic diagram showing the process of DNA replication in the exponential phase. 1) The origin replication complex binds to the *oriC* region. 2) The first round of replication is initiated when the first replication complex binds the *oriC* region. 3) The second round of replication begins before completion of the first round of replication. 4) Six replication forks are generated in one bacterium. 5) The first round of replication completes, followed by a new cell replication cycle. **D.** Fitting of the genomic coverages with the mathematical model of replication. The genomic coverage plots of 11492 (upper) and NTUH-K2044 (lower) strains are shown in gray, and the mathematically fitting curves are shown in red. t_D, doubling time; *oriC*, replication origin.

K. pneumoniae strains. Two t_D values were then calculated by fitting growth curves (~ 34.64 min for strain 11492; ~ 27.95 min for NTUH-K2044; Figure 5A). We could therefore infer t_R from the t_D/t_R (~ 58.71 min and ~ 47.36 min for strains 11492 and NTUH-K2044, respectively).

We further obtained the re-methylation time per motif (**Table 6**) by simulating the dynamic processes of passive demethylation and re-methylation in the exponential phase based on five parameters: the methylation read fraction of each motif ($M_{(x)}$), the initial methylation read fraction of each motif

Table 6 Average re-methylation time per motif in the NTUH-K2044 and 11492 strains

Motif sequence	No. of motifs (%)	Re-methylation time (min)
GATC	59,424 (100%)	3.52 ± 0.54
SGATCS (fast)	17,041 (28.7%)	3.41 ± 0.54**
WGATCW (slow)	9803 (16.5%)	3.78 ± 0.54***
CCWGG	35,516 (100%)	9.23 ± 4.93
SCCWGGS (slow)	13,573 (38.2%)	9.65 ± 4.88***
WCCWGGW (fast)	5800 (16.3%)	8.27 ± 4.48***
MTCGAK	10,467 (100%)	4.55 ± 0.54
GRACRAC	2002 (100%)	3.46 ± 0.45

Note: Degenerate bases used in our recognition sequences are as follows: R = G or A, Y = C or T, M = A or C, K = G or T, S = G or C, W = A or T, B = not A (C or G or T), D = not C (A or G or T), H = not G (A or C or T), V = not T (A or C or G), N = A or C or G or T. Values for re-methylation time are shown as mean ± SD; ** and *** means that the re-methylation time of the motifs with certain flanking sequences show the significant difference (**, $P < 0.01$; ***, $P < 0.001$) comparing with the average re-methylation time of total motifs.

($M^0_{(x)}$), t_D, t_R, and the distribution density of the first replication forks ($P_{(x_1)}$) (see Materials and methods for details).

In general, the mean re-methylation time of 6mA was shorter (3.52 min and 3.46 min for GATC and GRACRAC motifs, respectively) than that of 5mC (9.23 min and 4.55 min for CCWGG and MTCGAK motifs, respectively) in the exponential phase (Table 6). Regarding specific motifs, the re-methylation time in the NTUH-K2044 strain was slightly shorter than that in the 11492 strain (Figure S14). In addition, the flanking bases could influence the re-methylation rates of the GATC and CCWGG motifs; when the flanking bases were C/G rather than T/A, the GATC motif exhibited a faster re-methylation, which was completely reversed for the CCWGG motif (Table 6; Figure S15).

The re-methylation of GATC motifs in the intergenic and *oriC* regions is slow at the exponential phase

To investigate the role of GATC motifs in transcriptional regulation in *K. pneumoniae* strains, we analyzed the re-methylation rates of the motifs in IGRs (including 5′ USRs and 3′ DSRs) at the exponential phase (**Figure 6**A). The results indicated slower re-methylation of GATC motifs in IGRs (3.94 ± 5.82 min) than in GRs (3.50 ± 5.10 min). Since most 5′ USRs in bacteria overlap with the promoter regions, and are involved in transcriptional regulation [26], we further explored the COG functional categories of the top 5% of genes with the slowest re-methylation sites (> 7.04 min/motif) in 5′ USRs (Table S15). Four enriched functional categories (“cell cycle control, cell division, chromosome partitioning”, “carbohydrate transport and metabolism”, “intracellular trafficking, secretion, and vesicular transport”, and “translation, ribosomal structure and biogenesis”) were observed (Table S15), of which at least 5% of genes contained slow re-methylation sites (> 7.04 min/motif) in 5′ USRs. Among them, the “carbohydrate transport and metabolism” functional category accounted for the most genes (22) with slow re-methylation in 5′ USRs, including three hemi-/un-methylated motifs shared in the 14 *K. pneumoniae* strains, as well as one hemi-/un-methylated motif cluster.

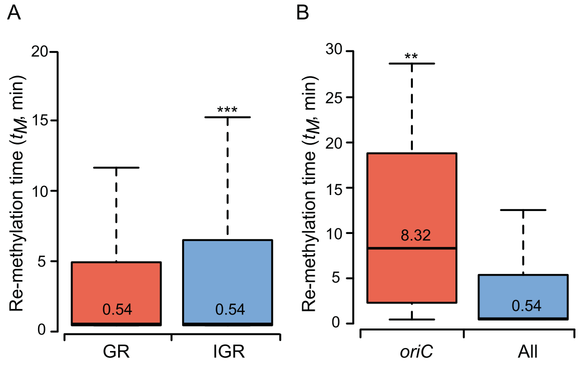

Figure 6 Comparison of re-methylation time of GATC motifs in different genome regions

A. Box plot showing the re-methylation time of GATC motifs in GRs and IGRs. **B.** Box plot showing the re-methylation time of GATC motifs in *oriC* and all regions. The value in each box represents the midian. **, $P < 0.01$; ***, $P < 0.001$. t_M, mean re-methylation time.

To further explore the role of the GATC motifs in regulating replication in *K. pneumoniae* strains, we analyzed the re-methylation rates of the GATC motifs in *oriC* regions (Figure 6B). The GATC motifs in *oriC* regions had the highest distribution density (~ 34 sites/kb) (Figure S8) and slower re-methylation rates (10.35 ± 8.69 min), compared to the average re-methylation rate of all genomic regions. Among the 14 GATC motifs in *oriC* regions, nine and eight motifs with slow re-methylation (> 7.04 min/motif) were enriched in the upstream sequence of the AT-rich region in 11492 and NTUH-K2044 strains, respectively (Figure S16).

Discussion

In this study, we precisely characterized the methylomes of 14 *K. pneumoniae* strains differing in serotypes, MLSTs, CGs, viscosity/virulence, and drug resistance using SMRT/bisulfite sequencing, and identified 15 DNA methylation motifs (eight of which were novel) corresponding to 13 R-M system MTases and two orphan MTases. Two motifs (GATC and CCWGG) and their respective orphan MTases (Dam and Dcm) appeared to be the most important since they were present in all *K. pneumoniae* strains, with more extensive distributions in genomes than other motifs (Tables 2 and 4). This feature was also reported previously for almost all members of the Enterobacteriaceae family [21,24]. Functional analysis in previous reports indicated that these two motifs, especially GATC, may perform multiple functions including transcriptional regulation, cell cycle control, and mismatch repair in *E. coli* and *S. enterica* [15–17,19–21]. We also demonstrated their high-density distributions in the *K. pneumoniae* genome relative to the simulated genome (Figure 3, Figures S9 and S10), presumably due to purifying selection (Figure S11), leading to their evolutionary conservation (Figure S12), as reported for other bacterial genomes [27]. This well-conserved feature also implies an essential role of these two motifs in *K. pneumoniae*.

The hemi-/un-methylated GATC motifs tended to be localised in IGRs, including 5′ USRs and 3′ DSRs (Figure 4A and C). Since promoters in *K. pneumoniae* were also predicted to be located throughout the 5′ USRs (Figure S17), these hemi-/un-methylated GATC motifs in 5′ USRs might be protected from methylation by competitively binding certain regulators to the promoter regions [17,19]. In *E. coli* strains, this feature facilitates the epigenetic regulation of downstream gene expression [28]. The status of these motifs suggests similar epigenetic mechanisms in *K. pneumoniae* strains, which might be a consequence of selection during long-term evolution (Figure S11), since our findings suggest that the genomic fragments with hemi-/un-methylated GATC motifs in IGRs have higher sequence conservation (Figure S12).

Importantly, by establishing a mathematical model to simulate the dynamic processes of passive demethylation and re-methylation for each motif in the exponential phase, we derived the re-methylation time for each motif throughout the whole genome (Table 6; Figure S14). Our studies revealed that the motifs at different genomic locations display different re-methylation rates. We could reasonably infer that the slower re-methylation in some sites/regions might also be due to the competitive binding of certain proteins to prevent methylation [17,19]. Thus, the slow re-methylation could precisely reflect the methylation-mediated epigenetic regulation at these sites/regions *in vivo*.

There are two types of methylation-mediated epigenetic regulation, transcription and replication, and both were examined by methylation kinetic analysis in our study. Firstly, the transcriptional regulation analysis indicated that the GATC motifs in IGRs present slower re-methylation than those in GRs (Figure 6). This is consistent with the distribution characteristics of the GATC motifs in IGRs (*i.e.*, lower methylated read ratios and more hemi-/un-methylated motifs; Figure 4A and B). As described above, most IGRs in bacteria overlap with the promoter regions and participate in transcriptional regulation [26]. Thus, the promoter regions with slow re-methylation motifs should be the locations where the GATC motifs with hemi-/un-methylated status function as transcription regulators in *K. pneumoniae* strains. Similarly, this should also be a consequence of the competitive binding between Dam and certain transcription regulatory proteins (such as OxyR) to these sites/regions [15,28,29], which is analogous to the epigenetic transcriptional regulation mediated by competition between DNA MTases and CCCTC-binding factor (CTCF) in CpG islands of eukaryotic cells [30]. Secondly, replication regulation analysis identified 7–8 slow GATC motifs enriched in the upstream sequence of the fourth DnaA binding site (DnaA box) and the adjacent AT-rich region (Figure S16). SeqA has been shown to bind preferentially to these motifs in order to lower re-methylation rates [31] and prevent the initiation cascade for chromosome replication [17,31]. Therefore, re-methylation of these motifs in the fourth DnaA box and AT-rich regions likely represents the main rate-limiting steps for triggering DNA replication initiation in *K. pneumoniae* strains. We also identified numerous promoter regions with slower re-methylation motifs in *K. pneumoniae* strains (Figure 6); therefore, it is reasonable to speculate that epigenetic regulation in bacteria is complex rather than simple, as previously believed.

Compared with the GATC motif, the CCWGG motif has different distribution characteristics in GRs and IGRs (Figure 4A and B). We therefore performed COG functional analysis of genes with hemi-/un-methylated CCWGG sites in IGRs, since previous studies demonstrated the epigenetic regulation of hemi-/un-methylated sites in IGRs in bacteria [32]. Our findings showed that the top three enriched functional categories for genes with hemi-/un-methylated CCWGG sites in IGRs were "replication, recombination and repair", "cell motility", and "coenzyme transport and metabolism" (Figure S18A), suggesting the possible epigenetic regulation of CCWGG methylation in *K. pneumoniae*. However, the analysis implied totally different COG categories for genes with hemi-/un-methylated GATC sites in IGRs (Figure S18B), suggesting differences in epigenetic regulation between the two main methylated sites in *K. pneumoniae*.

We also explored the methylation kinetic features of CCWGG motifs during the growth cycle. A total of 1961 CCWGG sites were differentially methylated during the growth cycle: 1954 sites that were hemi-/un-methylated in the exponential phase became methylated in the stationary phase, and seven sites that were un-methylated in the exponential phase became hemi-methylated in the stationary phase. No sites that were methylated in the exponential phase became hemi-/un-methylated in the stationary phase. Such dynamic changes in methylation status might be due to active and inactive replication in exponential and stationary phases [13]. Active replication at the exponential phase may result in more hemi-/un-methylated sites in the genome due to replication-mediated passive demethylation; inactive replication at the stationary phase may lead to more methylated sites in the genome due to almost complete MTase-catalyzed methylation. Incidentally, seven hemi-methylated CCWGG sites in the stationary phase are located in genes encoding four small subunit ribosomal RNAs, two mobile element proteins, and one possible transcription regulator (Table S16), which may be necessary for the survival of bacteria [33].

Importantly, eight novel MTases and related motifs were detected, including five Type I and three Type II MTases (Table 5, Tables S6 and S8). These novel MTases and cognate restriction endonucleases (REases) form R-M systems, which are known to protect bacterial cells by cleaving foreign phage DNA [9–14]. We subsequently investigated the hemi-/un-methylated sites in IGRs, since previous studies demonstrated epigenetic regulation of hemi-/un-methylated sites in IGRs of some bacteria [32]. Among the eight novel methylation mofits, the MTCGAK motifs were shared by the NTUH-K2044 and 11492 strains, which possessed the most hemi-/un-methylated sites (56.77% and 56.03% for NTUH-K2044 and 11492 strains, respectively) among the strains investigated (Table 4). More than 60% of un-methylated MTCGAK motifs were shared by the two strains (Table S17), which were mainly enriched in the "transcription", "inorganic ion transport and metabolism", and "replication, recombination and repair" COG categories (Figure S19). The CCAYN_7TTYG motifs were also present in two strains (23 and 11311) (Table 3). Three hemi-methylated CCAYN_7TTYG motifs were shared by these two strains, associated with *ascB*, *deoC*, and *fecA* genes, respectively (Table S18). These hemi-/un-methylated MT**C**$\underline{\text{G}}$AK and CC**A**YN$_7$$\underline{\text{T}}$TYG motifs in IGRs were shared by two strains, and they may contribute to the epigenetic regulation of related gene expression. In addition, we also performed COG analysis on the other six newly identified methylation motifs, which were only detected in one of the *K. pneumoniae* strains (Figure S19). These hemi-/un-methylated sites in only one *K. pneumoniae* strain may be derived from replication-mediated passive demethylation [34], which represents a limitation.

Most previous studies on MTase kinetics focused on *in vitro* analysis [35–37]. Several analyses on *in vivo* methylation kinetics have been reported, including one publication on bacteria, but have not explored single-motif methylation kinetics [38–40]. Our current study is the first to characterize the *in vivo* methylation kinetics at single motif resolution throughout the whole genome. The findings provide valuable resources to better understand epigenetic regulation in this and other bacterial species.

Materials and methods

Strain information, growth curves, and phenotypic characterisation

Details of the 14 *K. pneumoniae* strains used in this study are included in Table S1. The strains were cultured overnight in Luria-Bertani (LB) medium at 37 °C. A 1 ml sample of overnight culture was transferred to a flask containing 200 ml of LB medium and cultured in a shaker at 200 r/min. The bacterial growth curve was determined by recording OD_{600} values at different time points. This experiment was performed in triplicate.

Drug susceptibility tests were performed by VITEK 2 (bioMérieux, Durham, NC), and drug resistance phenotypes were determined by Clinical and Laboratory Standards Institute (CLSI) standards (https://clsi.org/standards/products/microbiology/documents/m100/). The string test was conducted on *K. pneumoniae* strains, and hypermucoviscosity was defined by viscous strings with a length of more than 5 mm [41].

Genomic DNA extraction, sequencing, assembly, correction, and annotation

Genomic DNA was extracted using TIANamp Bacteria Genomic DNA Kit (Catalog No. DP302, TIANGEN, Beijing, China). Whole-genome sequencing was performed using a PacBio RS II platform with P6/C4 chemistry (Pacific Biosciences, CA). Each strain was sequenced using 1–2 SMRT cells with genome coverage of more than 50× (Table S2).

De novo assembly of the genome was performed using Hierarchical Genome Assembly Process 3 (HGAP3) in the SMRT Portal (v2.2.0; https://smrt-analysis.readthedocs.io/en/latest/SMRT-Pipe-Reference-Guide-v2.2.0/). Gap closing was completed by PBJelly [42]. Based on BLAST results, genome circularisation was finished by manually removing overlapping contig regions.

To correct polymer errors, we re-sequenced strains using Illumina sequencing (Table S3). Paired-end libraries were prepared, and clean reads were obtained after eliminating redundant and low-quality raw reads. Paired reads were extracted and then mapped onto the assembled genome sequences to obtain unique mapped reads using BWA [43], and Pilon (v1.13) was subsequently employed to polish genome sequences using unique mapped reads [44].

Genome sequences were annotated by the Rapid Annotation using Subsystem Technology (RAST) [45]. Unannotated genes were then predicted by alignment in the NCBI non-redundant (NR) database using BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Protein functions were annotated based on COG, tRNAs were predicted by tRNAscan-SE [46], and rRNAs were predicted by 'search_for_rnas' tools in the RAST server (http://RAST.nmpdr.org).

Genome structure and phylogenetic analysis

ANI and coverage were calculated by ANI on EzBioCloud (http://www.ezbiocloud.net/tools/ani) and online BLAST. Multiple alignments of genomic sequences were performed using Mauve multiple alignment software [23].

Single nucleotide polymorphisms (SNPs) were detected by MUMmer [47] based on the 14 genomes in this study and 62 published genomes (Table S4) using HS11286 as reference. PRANK [48] was used to annotate the protein-coding genes of the 76 *K. pneumoniae* genomes, and Roary was employed to predict 3173 core genes [49]. SNPs (117,142) in core genes were used to construct a phylogenetic tree based on maximum likelihood using FastTree [50], followed by decoration using evolview (v2) [51].

Genome-wide detection of 6mA and related motifs using SMRT sequencing data

The SMRT Portal (v2.2.0) was applied to detect genome-wide 6mA modification and related motifs using standard settings in the "RS_Modification_and_Motif_Analysis.1" protocol as previously described (https://smrt-analysis.readthedocs.io/en/latest/SMRT-Pipe-Reference-Guide-v2.2.0/). We subsequently identified the motifs by selecting the top 1000 kinetic hits and submitting a window of ±20 bases around the detected modified base to MEME-ChIP [52], and then compared the results with predicted MTase targeting motif sequences in REBASE [24].

There are three methylation patterns for motifs: methylated, hemi-methylated, and un-methylated. Methylated/hemi-methylated/un-methylated motifs indicate sites with methylated nucleotides on both/one/no strands, respectively.

Genome-wide detection of 5mC and related motifs using bisulfite sequencing data

The 5mC methylation was detected by bisulfite sequencing (Table S5). Trimmomatic (v0.32) [53] was used to trim adapters and low-quality bases using default parameters. Clean reads were mapped against reference genomes by Bismark (v0.12.2) [54]. We identified motifs by submitting a window of ±20 bases around the detected modified base to MEME-ChIP [52], and then compared the results with predicted MTase targeting motif sequences in REBASE [24].

MTase cloning and verification

The predicted MTase genes were amplified from bacterial genomic DNA using gene-specific primers (Table S19) and cloned into the plasmid pRRS as previously described [14]. The corresponding methylation-sensitive restriction sites (used to detect the activity of MTases) were included in the 3′-end oligonucleotides. The recombinant plasmids were transformed into the ER2796 bacterial host (not containing known MTase genes), followed by bacterial culture overnight. Plasmids were prepared using QIAprep Spin Miniprep Kit (Catalog No. 27104, QIAGEN). The appropriate restriction enzymes were then used to determine the presence or absence of methylation motifs in recombinant plasmids. Digestion reactions were conducted for 4 h at 37 °C and products were separated on 1% agarose gels. Methylation motifs were further confirmed by SMRT/bisulfite sequencing of the recombinant plasmids.

Density distributions of the GATC and CCWGG motifs in *K. pneumoniae* and simulated genomes

The simulated genome was generated based on the same length and GC content as the *K. pneumoniae* genomes. We then calculated the number of motifs in consecutive 1-kb segments to determine the density of GATC and CCWGG motifs across the *K. pneumoniae* and simulated genomes.

We defined high-/low-density regions by calculating the number of motifs in each 2-kb non-overlapping sliding window in the genome. Through normal distribution analysis using the "pnorm" function in R, the top and bottom 5% of regions were defined as high-density and low-density regions, respectively.

Ka/Ks analysis of the GATC and CCWGG motif sequences and corresponding scramble sequences

We first extracted the minimum DNA sequences (2–3 codons) containing the motifs from the open reading frames of genes to obtain motif sequences. Corresponding scramble sequences were obtained by random shuffling (eliminating the motif sequences). The top 10 most frequent scramble sequences were used in subsequent analysis. We then identified the reference sequences of motif sequences/scramble sequences in the reference genome (HS11286) by Multiple Alignment using Fast Fourier Transform (MAFFT) [55]. The motif sequences, scramble sequences, and corresponding reference sequences of each strain were respectively concatenated, and their Ka/Ks ratios were calculated by ParaAT [56].

Sequence conservation analysis of the GATC and CCWGG motifs

We extracted the methylated motifs and their flanking sequences (20 nt) in IGRs from one genome and obtained the corresponding sequences of 13 other genomes by multiple sequence alignment. Using the aforementioned 14 sequences, conservation scores were calculated by PhyloP [57]. The conservation scores of hemi-/un-methylated motifs with 20 nt flanking sequences in IGRs were calculated using the same method.

Calculation of genomic replication termination time through simulating genomic coverage plots

Genomic coverage plots reflect the accumulated copy numbers across genomes. Since replication forks always proceed from the *oriC* to the doubling point (s_{DP}), only a few can get close to the doubling point in the exponential phase. As a result, the sequencing coverage in the *oriC* region is much higher than that in the doubling point region. In each cell, the copy number for each site (s) depends on its relative position to the first replication fork (s_1) and the minimal time of successive initiations (re-initiation). The re-initiation time is consistent with the t_D in the exponential phase, reflecting the growth rate of *K. pneumoniae* strains.

After initiation, replication forks should advance bidirectionally at similar speeds in the exponential phase. Therefore, genome coverage plots are symmetrical near the doubling point. We divided the coverage plots into left ($s \leq s_{DP}$) and right ($s \geq s_{DP}$) parts by doubling point, and simulated the curves independently. We then obtained the relative position ($x \in (0,1]$) of each genomic site using Equation (1) where S_G represents the half-length of the genome.

$$\begin{cases} x = s/S_G & (s \leq s_{DP}) \\ x = 2 - s/S_G & (s \geq s_{DP}) \end{cases} \tag{1}$$

We subsequently calculated the relative distance of the first replication fork when another one at *oriC* is re-initiated (Δx_R) using Equation (2) where t_R represents the genomic replication termination time.

$$\Delta x_R = t_D/t_R \ (0 \leq t_D \leq t_R) \tag{2}$$

By constructing the replication frequency matrix of genomic sites (x) with different distances from the first replication fork (x_1), the copy number of the genomic site ($f_{(x,x_1)}$) can be determined using Equation (3). When the genomic site is in front of the first replication fork ($x \geq x_1$), its copy number is 1; when the site is between the first and second replication forks ($x_1 - \Delta x_R \leq x \leq x_1$), its copy number is 2. The rest were deduced by analogy.

$$f_{(x,x_1)} = \left\lceil 2^{\lceil (x_1 - x)/\Delta x_R \rceil} \right\rceil \tag{3}$$

Next, we used the B distribution to assess cell population density with different genomic positions of the first replication forks ($x_1 \in [0,1]$). Since x is not continuous data (step size = $1/S_G$), the cell population density of the first replication forks ($P_{(x_1)}$) at each genomic site was determined from the difference of accumulated densities ($I_x(\alpha,\beta)$) of the adjacent sites.

$$P_{(x_1)} = I_{(x_1)}(\alpha,\beta) - I_{\left(x_1 - \frac{1}{S_G}\right)}(\alpha,\beta) \tag{4}$$

Genome coverage plots could be determined as the integral of the copy numbers of each genomic site in the cellular populations ($H(x)$) using Equation (5). The integral was calculated based on $P_{(x_1)}$ and the relative distance between x and x_1 ($x_1 - x$). By substituting Equations (2)–(4) into Equation (5), we established a mathematical model to fit the genome coverage plots using Equation (6).

$$H(x) = \int_0^1 f_{(x,x_1)} * P_{(x_1)} d_{x_1} \tag{5}$$

$$H(x) = \int_0^1 \lceil 2^{\lceil (x_1 - x) * t_R / t_D \rceil} \rceil * \left(I_{(x_1)}(\alpha, \beta) - I_{\left(x_1 - \frac{1}{S_G}\right)}(\alpha, \beta) \right) d_{x_1} \tag{6}$$

We then substituted Equation (1) into Equation (6), and repeatedly fitted the genome coverage plots by selecting continuous parameters with step sizes of 0.1. We finally obtained the optimal solutions of Δx_R and B distribution parameters (α,β) through goodness of fit tests. Since t_D could be calculated from the growth curve of each *K. pneumoniae* strain, we could obtain the t_R using Equation (2).

Calculation of the re-modification time of motifs

In the exponential phase, the methylation read fraction ($M_{(x)}$) of each motif was determined from its initial methylation read fraction ($M^0_{(x)}$), replication-induced passive demethylation, and MTase-catalyzed re-methylation. Based on $M_{(x)}$ in the exponential phase and B distribution parameters (α,β), we calculated the proceeding distance of the corresponding replication fork when the hemi-methylated motif was re-methylated (Δx_M). As shown in Equation (7), Δx_M corresponds to the ratio of mean re-methylation time (t_M) to the genomic replication termination time (t_R).

$$\Delta x_M = t_M / t_R \tag{7}$$

When motif x is located downstream of a replication fork ($x_j \leq x$), its methylation read fraction ($M_{(x)}$) remains unchanged due to no replication-induced passive demethylation. When motif x is located upstream of a replication fork with a longer distance ($x_j \geq x + \Delta x_M$), the methylation read fraction is also unchanged due to the completion of re-methylation of motif x. Thus, we only considered the demethylation effect of replication forks in a certain range ($x \leq x_j \leq (x + \Delta x_M)$) on $M_{(x)}$ in the mathematical derivation, where j represents the serial number of the replication fork affecting $M_{(x)}$.

If the hemi-methylated motifs can be rapidly re-modified before the next replication fork ($t_M \leq t_D$), we only need to evaluate the demethylation effect of one replication for all replication forks in the range of $x \leq x_j \leq x + \Delta x_M$. Here, we include the first replication forks (x_1) and other effective replication forks (x_j) in the range. We then substituted the accumulated B distribution density of the first replication forks $(x + (j-1)\Delta x_R \leq x_1 \leq x + \Delta x_M + (j-1)\Delta x_R)$ with that of the corresponding effective replication forks $(x \leq x_j \leq x + \Delta x_M)$. Due to semi-conservative replication, we further obtain Δx_M from Equation (8) where N represents the upper limit of j, which is determined by Δx_R and the relative genomic position (x) of methylated motifs.

$$\begin{cases} M(x) = M^0_{(x)} * \left(1 - 0.5 * \sum_{j=1}^{N} \begin{pmatrix} I_{(x + \Delta x_M + (j-1) * \Delta x_R)}(\alpha, \beta) - \\ I_{(x + (j-1) * \Delta x_R)}(\alpha, \beta) \end{pmatrix} \right) \\ N = \lceil (1 - x) / \Delta x_R \rceil \end{cases} \tag{8}$$

If the methylation is slow, the hemi-methylated motifs may not be re-modified when the next replication fork crosses them ($t_M > t_D$). In this case, the effect of multiple passive demethylation on $M_{(x)}$ should be considered using Equation (9), which is determined by the ratio of mean re-methylation time (t_M) to re-initiation time (t_D), where n represents the upper limit of multiple passive demethylation times (i).

$$n = \lceil t_M / t_D \rceil = \lceil \Delta x_M / \Delta x_R \rceil \tag{9}$$

We then converted the genomic positions of effective replication forks (x_j) into those of the corresponding first replication forks (x_1), and substituted the accumulated B distribution density of x_1 $(x + (j-1) * \Delta x_R \leq x_1 \leq x + j * \Delta x_R)$ for the x_j population density $(x \leq x_j \leq x + \Delta x_M)$. Considering the differential influences on $M_{(x)}$ in the case of $j < n$ and $n \leq j \leq N$, we calculated Δx_M using Equation (10) by comprehensive calculations.

$$\begin{cases} M(x) = M^0_{(x)} * \begin{pmatrix} 1 - \sum_{j=1}^{n-1} \sum_{i=1}^{j} 0.5^i * \begin{pmatrix} I_{(x + j * \Delta x_R)}(\alpha, \beta) - \\ I_{(x + (j-1) * \Delta x_R)}(\alpha, \beta) \end{pmatrix} \\ - \sum_{j=n}^{N} \sum_{i=1}^{n-1} 0.5^i * \begin{pmatrix} I_{(x + j * \Delta x_R)}(\alpha, \beta) - \\ I_{(x + (j-1) * \Delta x_R)}(\alpha, \beta) \end{pmatrix} \\ -0.5^n * \sum_{j=n}^{N} \begin{pmatrix} I_{(x + \Delta x_M + (j-n) * \Delta x_R)}(\alpha, \beta) - \\ I_{(x + (j-1) * \Delta x_R)}(\alpha, \beta) \end{pmatrix} \end{pmatrix} \\ n = \lceil t_M / t_D \rceil = \lceil \Delta x_M / \Delta x_R \rceil \\ N = \lceil (1 - x) / \Delta x_R \rceil \end{cases} \tag{10}$$

Based on the growth curves and genome coverage plots of the sequencing data, we deduced various parameters in different *K. pneumoniae* strains, including t_D, t_R, and $I_x(\alpha, \beta)$. The mean re-methylation time (t_M) of each motif could be further calculated by substituting the aforementioned parameters, Equations (2) and (7), the methylation read fraction in the exponential phase ($M_{(x)}$), and the stationary phase ($M^0_{(x)}$) into Equation (10).

Data availability

The genome data of 14 *K. pneumoniae* strains have been deposited in the Genome Sequence Archive [58] at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences/ China National Center for Bioinformation (GSA: CRA003482), and are publicly accessible at https://bigd.big.ac.cn/gsa. The data are also at the NCBI database (BioProject: PRJNA477755), and are publicly accessible at https://www.ncbi.nlm.nih.gov.

CRediT author statement

Jing Fu: Validation, Writing - original draft, Resources. **Ju Zhang:** Methodology, Writing - original draft, Resources. **Li Yang:** Software, Formal analysis, Writing - original draft. **Nan Ding:** Formal analysis, Writing - original draft, Visualization. **Liya Yue:** Validation, Writing - original draft, Visualization. **Xiangli Zhang:** Formal analysis, Visualization. **Dandan Lu:** Formal analysis, Visualization. **Xinmiao Jia:** Visualization. **Cuidan Li:** Visualization, Data curation. **Chongye Guo:** Visualization. **Zhe Yin:** Resources. **Xiaoyuan Jiang:** Resources. **Yongliang Zhao:** Writing - review & editing. **Fei Chen:** Conceptualization, Writing - review & editing, Funding acquisition, Project administration, Supervision. **Dongsheng Zhou:** Conceptualization, Writing - review & editing, Project administration, Supervision. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by the National Key Program for Infectious Disease of China (Grant No. 2018ZX10302301-004-003) and the National Natural Science Foundation of China (Grant Nos. 31770870 and 31801093).

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2021.04.002.

ORCID

ORCID 0000-0001-7961-0888 (Jing Fu)
ORCID 0000-0003-3274-9844 (Ju Zhang)
ORCID 0000-0001-7353-4415 (Li Yang)
ORCID 0000-0002-1045-1695 (Nan Ding)
ORCID 0000-0002-5752-6914 (Liya Yue)
ORCID 0000-0003-4816-6598 (Xiangli Zhang)
ORCID 0000-0001-6977-4301 (Dandan Lu)
ORCID 0000-0002-6084-0146 (Xinmiao Jia)
ORCID 0000-0002-8430-1155 (Cuidan Li)
ORCID 0000-0002-1918-8643 (Chongye Guo)
ORCID 0000-0003-0530-7100 (Zhe Yin)
ORCID 0000-0002-8962-7170 (Xiaoyuan Jiang)
ORCID 0000-0003-0121-1312 (Yongliang Zhao)
ORCID 0000-0003-3895-6318 (Fei Chen)
ORCID 0000-0003-3827-5969 (Dongsheng Zhou)

References

[1] Guo Y, Wang S, Zhan L, Jin Y, Duan J, Hao Z, et al. Microbiological and clinical characteristics of hypermucoviscous *Klebsiella pneumoniae* isolates associated with invasive infections in China. Front Cell Infect Microbiol 2017;7:24.

[2] Lam MMC, Wyres KL, Duchene S, Wick RR, Judd LM, Gan YH, et al. Population genomics of hypervirulent *Klebsiella pneumoniae* clonal-group 23 reveals early emergence and rapid global dissemination. Nat Commun 2018;9:2703.

[3] Fang CT, Lai SY, Yi WC, Hsueh PR, Liu KL, Chang SC, et al. *Klebsiella pneumoniae* genotype K1: an emerging pathogen that causes septic ocular or central nervous system complications from pyogenic liver abscess. Clin Infect Dis 2007;45:284–93.

[4] Bialek-Davenet S, Criscuolo A, Ailloud F, Passet V, Jones L, Delannoy-Vieillard AS, et al. Genomic definition of hypervirulent and multidrug-resistant *Klebsiella pneumoniae* clonal groups. Emerg Infect Dis 2014;20:1812–20.

[5] Hammond AW, Gerard GF, Chatterjee DK. Cloning the *Kpn*I restriction-modification system in *Escherichia coli*. Gene 1991;97:97–102.

[6] Valinluck B, Lee NS, Ryu J. A new restriction-modification system, *Kpn*BI, recognized in *Klebsiella pneumoniae*. Gene 1995;167:59–62.

[7] Mehling JS, Lavender H, Clegg S. A Dam methylation mutant of *Klebsiella pneumoniae* is partially attenuated. FEMS Microbiol Lett 2007;268:187–93.

[8] Fang CT, Yi WC, Shun CT, Tsai SF. DNA adenine methylation modulates pathogenicity of *Klebsiella pneumoniae* genotype K1. J Microbiol Immunol Infect 2017;50:471–7.

[9] Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, et al. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. Nat Biotechnol 2012;30:1232–9.

[10] Bottacini F, Morrissey R, Roberts RJ, James K, van Breen J, Egan M, et al. Comparative genome and methylome analysis reveals restriction/modification system diversity in the gut commensal *Bifidobacterium breve*. Nucleic Acids Res 2018;46:1860–77.

[11] Lluch-Senar M, Luong K, Llorens-Rico V, Delgado J, Fang G, Spittle K, et al. Comprehensive methylome characterization of *Mycoplasma genitalium* and *Mycoplasma pneumoniae* at single-base resolution. PLoS Genet 2013;9:e1003191.

[12] Hargreaves KR, Thanki AM, Jose BR, Oggioni MR, Clokie MRJ. Use of single molecule sequencing for comparative genomics of an environmental and a clinical isolate of *Clostridium difficile* ribotype 078. BMC Genomics 2016;17:1020.

[13] Murray IA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, et al. The methylomes of six bacteria. Nucleic Acids Res 2012;40:11450–62.

[14] Zhu L, Zhong J, Jia X, Liu G, Kang Y, Dong M, et al. Precision methylome characterization of *Mycobacterium tuberculosis* complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. Nucleic Acids Res 2016;44:730–43.

[15] Cota I, Bunk B, Sproer C, Overmann J, Konig C, Casadesus J. OxyR-dependent formation of DNA methylation patterns in OpvABOFF and OpvABON cell lineages of *Salmonella enterica*. Nucleic Acids Res 2016;44:3595–609.

[16] Cohen NR, Ross CA, Jain S, Shapiro RS, Gutierrez A, Belenky P, et al. A role for the bacterial GATC methylome in antibiotic stress survival. Nat Genet 2016;48:581–6.

[17] Wolanski M, Donczew R, Zawilak-Pawlik A, Zakrzewska-Czerwinska J. *oriC*-encoded instructions for the initiation of bacterial chromosome replication. Front Microbiol 2015;5:735.

[18] Kozdon JB, Melfi MD, Luong K, Clark TA, Boitano M, Wang S, et al. Global methylation state at base-pair resolution of the *Caulobacter* genome throughout the cell cycle. Proc Natl Acad Sci U S A 2013;110:E4658–67.

[19] Lobner-Olesen A, Hansen FG, Rasmussen KV, Martin B, Kuempel PL. The initiation cascade for chromosome replication in wild-type and Dam methyltransferase deficient *Escherichia coli* cells. EMBO J 1994;13:1856–62.

[20] Lahue RS, Su SS, Modrich P. Requirement for d(GATC) sequences in *Escherichia coli* mutHLS mismatch correction. Proc Natl Acad Sci U S A 1987;84:1482–6.

[21] Blow MJ, Clark TA, Daum CG, Deutschbauer AM, Fomenkov A, Fries R, et al. The epigenomic landscape of prokaryotes. PLoS Genet 2016;12:e1005854.

[22] Lee CR, Lee JH, Park KS, Kim YB, Jeong BC, Lee SH. Global dissemination of carbapenemase-producing *Klebsiella pneumoniae*: epidemiology, genetic context, treatment options, and detection methods. Front Microbiol 2016;7:895.

[23] Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 2010;5:e11147.

[24] Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE — a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res 2015;43:D298–9.

[25] Hurst LD. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 2002;18:486.

[26] de Jong A, Pietersma H, Cordes M, Kuipers OP, Kok J. PePPER: a webserver for prediction of prokaryote promoter elements and regulons. BMC Genomics 2012;13:299.

[27] Luo H, Gao F, Lin Y. Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. Sci Rep 2015;5:13210.

[28] Waldron DE, Owen P, Dorman CJ. Competitive interaction of the OxyR DNA-binding protein and the Dam methylase at the antigen 43 gene regulatory region in *Escherichia coli*. Mol Microbiol 2002;44:509–20.

[29] Campbell JL, Kleckner N. *E. coli oriC* and the *dnaA* gene promoter are sequestered from *dam* methyltransferase following the passage of the chromosomal replication fork. Cell 1990;62:967–79.

[30] Chang SW, Chao WR, Ruan A, Wang PH, Lin JC, Han CP. A promising hypothesis of c-KIT methylation/expression paradox in c-KIT (+) squamous cell carcinoma of uterine cervix — CTCF transcriptional repressor regulates c-KIT proto-oncogene expression. Diagn Pathol 2015;10:207.

[31] Nievera C, Torgue JJ, Grimwade JE, Leonard AC. SeqA blocking of DnaA-*oriC* interactions ensures staged assembly of the *E. coli* pre-RC. Mol Cell 2006;24:581–92.

[32] Camacho EM, Casadesus J. Regulation of *traJ* transcription in the *Salmonella* virulence plasmid by strand-specific DNA adenine hemimethylation. Mol Microbiol 2005;57:1700–18.

[33] Lopez Sanchez MIG, Cipullo M, Gopalakrishna S, Khawaja A, Rorbach J. Methylation of ribosomal RNA: a mitochondrial perspective. Front Genet 2020;11:761.

[34] Radman-Livaja M, Liu CL, Friedman N, Schreiber SL, Rando OJ. Replication and active demethylation represent partially overlapping mechanisms for erasure of H3K4me3 in budding yeast. PLoS Genet 2010;6:e1000837.

[35] Coffin SR, Reich NO. Modulation of *Escherichia coli* DNA methyltransferase activity by biologically derived GATC-flanking sequences. J Biol Chem 2008;283:20106–16.

[36] Pollak AJ, Chin AT, Brown FL, Reich NO. DNA looping provides for "intersegmental hopping" by proteins: a mechanism for long-range site localization. J Mol Biol 2014;426:3539–52.

[37] Barel I, Naughton B, Reich NO, Brown FLH. Specificity versus processivity in the sequential modification of DNA: a study of DNA adenine methyltransferase. J Phys Chem B 2018;122:1112–20.

[38] Zheng Y, Sweet SM, Popovic R, Martinez-Garcia E, Tipton JD, Thomas PM, et al. Total kinetic analysis reveals how combinatorial methylation patterns are established on lysines 27 and 36 of histone H3. Proc Natl Acad Sci U S A 2012;109:13549–54.

[39] Wodarz D, Boland CR, Goel A, Komarova NL. Methylation kinetics and CpG-island methylator phenotype status in colorectal cancer cell lines. Biol Direct 2013;8:14.

[40] Hojfeldt JW, Laugesen A, Willumsen BM, Damhofer H, Hedehus L, Tvardovskiy A, et al. Accurate H3K27 methylation can be established *de novo* by SUZ12-directed PRC2. Nat Struct Mol Biol 2018;25:225–32.

[41] Siu LK, Yeh KM, Lin JC, Fung CP, Chang FY. *Klebsiella pneumoniae* liver abscess: a new invasive syndrome. The Lancet Infect Dis 2012;12:881–7.

[42] English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 2012;7:e47768.

[43] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.

[44] Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014;9:e112963.

[45] Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, et al. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep 2015;5:8365.

[46] Chan PP, Lowe TM. tRNAscan-SE: searching for tRNA genes in genomic sequences. Methods Mol Biol 2019;1962:1–14.

[47] Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol 2018;14:e1005944.

[48] Loytynoja A, Goldman N. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics 2010;11:579.

[49] Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015;31:3691–3.

[50] Price MN, Dehal PS, Arkin AP. FastTree 2 — approximately maximum-likelihood trees for large alignments. PLoS One 2010;5: e9490.

[51] He Z, Zhang H, Gao S, Lercher MJ, Chen WH, Hu S. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic Acids Res 2016;44: W236–41.

[52] Ma W, Noble WS, Bailey TL. Motif-based analysis of large nucleotide data sets using MEME-ChIP. Nat Protoc 2014;9:1428–50.

[53] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20.

[54] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 2011;27:1571–2.

[55] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;30:772–80.

[56] Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, et al. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. Biochem Biophys Res Commun 2012;419:779–81.

[57] Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 2005;15:901–13.

[58] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics 2021;19:578–83.