Unlocking the mystery of the hard-to-sequence phage genome: PaP1 methylome and bacterial immunity

Shuguang Lu¹, Shuai Le¹, Yinling Tan¹, Ming Li¹, Chang Liu², Kebin Zhang¹, Jianjun Huang¹, Haimei Chen², Xiancai Rao¹, Junmin Zhu¹, Lingyun Zou¹, Qingshan Ni¹, Shu Li¹, Jing Wang¹, Xiaolin Jin¹, Qiwén Hu¹, Xinyue Yao¹, Xia Zhao¹, Lin Zhang¹, Guangtao Huang¹ and Fuquan Hu¹*

Abstract
Background: Whole-genome sequencing is an important method to understand the genetic information, gene function, biological characteristics and survival mechanisms of organisms. Sequencing large genomes is very simple at present. However, we encountered a hard-to-sequence genome of *Pseudomonas aeruginosa* phage PaP1. Shotgun sequencing method failed to complete the sequence of this genome.

Results: After persevering for 10 years and going over three generations of sequencing techniques, we successfully completed the sequence of the PaP1 genome with a length of 91,715 bp. Single-molecule real-time sequencing results revealed that this genome contains 51 N-6-methyladenines and 152 N-4-methylcytosines. Three significant modified sequence motifs were predicted, but not all of the sites found in the genome were methylated in these motifs. Further investigations revealed a novel immune mechanism of bacteria, in which host bacteria can recognise and repel modified bases containing inserts in a large scale. This mechanism could be accounted for the failure of the shotgun method in PaP1 genome sequencing. This problem was resolved using the *nfi*⁻ mutant of *Escherichia coli* DH5α as a host bacterium to construct a shotgun library.

Conclusions: This work provided insights into the hard-to-sequence phage PaP1 genome and discovered a new mechanism of bacterial immunity. The methylome of phage PaP1 is responsible for the failure of shotgun sequencing and for bacterial immunity mediated by enzyme Endo V activity; this methylome also provides a valuable resource for future studies on PaP1 genome replication and modification, as well as on gene regulation and host interaction.

Background
Whole-genome sequencing is a very important method to understand the genotype and phenotype of an organism. In 1976, the genome of phage MS2 (only 3.5 kb in length) was the first completely sequenced genome [1]. The whole genome sequence of phage φX174 (with 5.3 kb genome) was then reported a year later [2]. Early genome-sequencing studies mainly focused on small genomes. With the advancement of sequencing technologies, particularly shotgun sequencing method [3,4], the sequencing of large genomes has become possible. Thus far, next- and third-generation sequencing technologies have become available [5-8]. Hence, genome sequencing has shown remarkable development.

However, small genomes, particularly bacteriophage genomes, are occasionally hard to be sequenced. We once encountered a tough work in sequencing a phage genome with a size of approximately 90 kb. In 2004, we isolated and characterised a *Pseudomonas aeruginosa* phage named PaP1 [9,10]. Pulsed-field gel electrophoresis (PFGE) results showed that PaP1 contains a genome of approximately 90 kb, but 20 contigs obtained using the shotgun library sequencing method could not be assembled in an integral genome; the total length of these obtained contigs was approximately 47.7 kb, which is almost half of 90 kb. We subsequently submitted the PaP1 genomic DNA to another sequencing center, where this DNA was subjected to repeated sequencing with the shotgun method. We obtained
almost the same result. We further verified this result by obtaining the PaP1 genome sequence with primer walking [11]; however, we failed again. Hence, this work was suspended.

Four years later, Roche/454 technique [12,13], a second-generation sequencing method, was established. We re-sequenced the PaP1 genome by using the Roche/454 technique in 2008. We easily obtained the complete PaP1 genome sequence with a size of 91,715 bp. Thus, we aimed to determine why the PaP1 genome was successfully sequenced using the Roche/454 DNA sequencer but not using the shotgun sequencing method. Based on the differences of the principles of the two sequencing methods, our presumption was that the host bacterium of the shotgun library construction, Escherichia coli DH5α, may greatly repel the inserted phage-DNA fragments by a particular immune mechanism. In the present study, this hypothesis was confirmed by conducting several experiments, including gene knockout and single-molecule real-time (SMRT) DNA sequencing techniques (third-generation sequencing methods) [6,14-16]; we also investigated the methylome of PaP1 (these particles were placed in an ice bath for 1 h and centrifuged at 12,000 × g for 10 min; the precipitate was then collected) and further purified using a CsCl gradient ultracentrifuge in accordance with previously reported methods [21,22].

DNA extraction and purification
EDTA (20 mM), proteinase K (50 μg mL⁻¹) and sodium dodecyl sulfate (0.5%, w/v) were added to the purified phage stock solution (PaP1 or PaP3). The mixture was incubated at 56°C for 1 h; an equal volume of phenol-chloroform-isoamyl alcohol solution (25:24:1) was added and the resulting mixture was centrifuged at 5,000 × g for 10 min. An aqueous layer was collected and extracted with chloroform at 5,000 × g for 10 min. The collected aqueous layer was mixed with 0.6 volumes of isopropanol and stored overnight at −20°C. Afterward, the mixture was centrifuged for 10 min at 12,000 × g and 4°C; the precipitated DNA was collected and washed with 70% and 100% ethanol, respectively. The PaP1 DNA was suspended in TE buffer (pH 8.0) and stored at −20°C for subsequent use.

Endonuclease digestion assay
The following restriction endonucleases were used to digest the genomic DNA of PaP1 or PaP3 in 20 μL reaction systems according to the manufacturer’s instructions: Paul; VspI; AatII; Spel; and EcoRI (New England Biolabs, Ipswich, MA, USA). The mixture was incubated at 37°C for 120 min and then used to perform PFGE. PFGE was conducted in 1% agarose gel with an initial switch time of 0.6 s and a final switch time of 1.6 s at 8 V/cm and an angle of 180° with a run time of 4.5 h. The restriction map was captured and analysed using Quantity One software (Bio-Rad, Hercules, CA, USA) to estimate the sizes of DNA bands on the gel. The commercial Endo V, or the products of E. coli gene nfi, was purchased from New England Biolabs, Ipswich, MA, USA. The PaP1 or PaP3 genomic DNA was digested by Endo V in 20 μL reaction systems according to the manufacturer’s instructions.

Sequencing of the PaP1 genome by using shotgun library method
In 2004, the genomic DNA of PaP1 was submitted to Chinese National Human Genome Center (CNHGC) in Shanghai, China for genome sequencing with the shotgun sequencing method [3] in an ABI 3730 DNA sequencer (ABI, Foster City, CA, USA). A shotgun library was constructed using E. coli DH5α as host bacterium. The PaP1 genomic DNA was digested by Sau3AI (New England Biolabs, Ipswich, MA, USA) or treated with ultrasonic waves; the DNA fragments with a length ranging from 1.6 kb to 2.0 kb were recovered to construct the shotgun library. The recovered DNA fragments were ligated into pUC18 and then electrotransformed into the host bacterium E. coli DH5α. Clones were
selected randomly from the library and used for sequencing. A total of 1,653 clones were sequenced and the average sequence coverage reached approximately 15-fold of the PaP1 genome. The obtained reads were assembled using the Phred/Phrap/Consed software package [23]. We obtained 20 contigs, but these contigs could not be assembled into an integral genome. To obviate mistakes caused by sequencing, we submitted the PaP1 genomic DNA to CNHGC in Beijing, China for repeat sequencing. Although the average sequence coverage also reached approximately 15-fold of the PaP1 genome, the obtained results were almost the same as those of the first sequencing. We also tried primer walking [11] to fill the gaps, but we failed to obtain the whole genome sequence of PaP1.

In 2012, we knocked out the nfi gene of E. coli DH5α (see below). To validate whether or not the nfi mutaant of E. coli DH5α can be used to construct a shotgun library and sequence the PaP1 genome, we repeated the sequencing of the PaP1 genome at Genemine Biotechnology Co., Ltd. (Chongqing, China). The procedures were exactly the same as described previously except the shotgun library clones were constructed with the nfi mutant of E. coli DH5α as host bacterium. At this time, 1,017 clones were sequenced and the average sequence coverage reached approximately 10-fold of the PaP1 genome.

### Sequencing of the PaP1 genome by using Roche/454 technique

In 2008, next-generation sequencing techniques were established. We then submitted the PaP1 genome to the CNHGC (Shanghai, China) for sequencing with a Roche/454 GS FLX titanium system [12]. In brief, the purified genomic DNA of PaP1 was fragmented, ligated to adapters and separated into single strands; the DNA fragments were bound to beads and amplified by emulsion PCR. A solid-phase pyrophosphate sequencing reaction was performed to reveal the raw sequence data. The Roche/454 reads were assembled using a Newbler assembler [24] (454 Life Sciences). The PaP1 genome sequence and its annotation information were available for download at the NCBI GenBank (http://www.ncbi.nlm.nih.gov/genbank/) with an accession number of HQ832595.

### Construction of the nfi mutant of E. coli DH5α

The nfi mutant of E. coli DH5α was constructed in accordance with previously described protocols [25,26]. The plasmids used in the procedure are listed in Table 1. The primers and other DNA sequences used in this procedure are listed in Table 2. The primers Cm-F [containing 55 bp upstream homologous extensions of the nfi gene (H1)] and Cm-R [containing 55 bp downstream homologous extensions of the nfi gene (H2)] were designed using the DNA sequence of pKD3 as a template. The PCR product

---

**Table 1 Bacterial strains and plasmids used in this study**

| Strains or plasmids | Relevant characteristics* | Source or reference |
|---------------------|---------------------------|---------------------|
| **Strains**         |                           |                     |
| Pseudomonas         | Belongs to serum typing 9  | Laboratory          |
| aeruginosa PA1      | of P. aeruginosa           | collection          |
|                     | international antigenic    |                     |
|                     | typing system; host of PaP1|                     |
| P. aeruginosa PA3   | Belongs to serum typing 6  | Laboratory          |
|                     | of P. aeruginosa           | collection          |
|                     | international antigenic    |                     |
|                     | typing system; host of PaP3|                     |
| E. coli DH5α        | Host for the construction  | Promega, WI, USA    |
|                     | of shotgun library clones  |                     |
| E. coli DH5α        | The nfi gene is replaced   | This study          |
| catΔnfi             | with cat.                  |                     |
| E. coli DH5αΔnfi    | The nfi gene is knocked    | This study          |
|                     | out.                      |                     |
| **Plasmids**        |                           |                     |
| pKD3                | Template plasmid for Red  | [26]                |
|                     | system; Amp’, Cm’         |                     |
| pKD46               | Red expression plasmid;    | [26]                |
|                     | Amp’                      |                     |
| pCP20               | Fli expression plasmid;    | [25]                |
|                     | Amp’, Cm’                 |                     |
| pUC18               | Vector for the construction| TaKaRa, Shiga, Japan|
|                     | of shotgun library clones;|                     |
|                     | Amp’                      |                     |

*Ampr, ampicillin resistant; Cmr, chloramphenicol resistant.

---

**Table 2 Primers and other DNA sequences used in this study**

| Primers or other DNA sequences* | Sequence (S’-3’) | Target genes or locations |
|---------------------------------|------------------|---------------------------|
| **Construction of the nfi mutant** |                  |                           |
| Cm-F                            | GTTGAGGCTGAGCTGCTTCC | Chloromycetin-resistant gene of pKD3 |
| Cm-R                            | CATATGAAATATCCCTTCTTAG |                           |
| Nfi-F                           | TGGCGCCGAGAAACATGC | nfi gene of E. coli DH5α  |
| Nfi-R                           | GACGCAGATGATTTGCTT |                           |
| H1                              | CGTGAGGCGGTATCGGCTGATACTGGATCAAGGCTGTTTACCGTGAATTGAT | Upstream of the nfi gene |
| H2                              | TTTTGAACATGTTGATTTACCGTTTACCGTGAATTGAAT | Downstream of the nfi gene |

*Primers and other DNA sequences were synthesised by BGI-Shenzhen (Shenzhen, China).
(donor DNA) that contains the chloramphenicol resistance gene (cat) and two FLP (a yeast-derived recombinase) recognition target (FRT) sites were then obtained by two-step PCR with Cm-F and Cm-R primers. The pKD46 plasmid (containing λ-Red recombinase) and the donor DNA were electrotransformed into E. coli DH5α. The bacteria were cultured in LB medium containing 100 mM L-arabinose (Sigma-Aldrich, St. Louis, MO) at 30°C for 12 h to induce homologous recombination between cat and nfi genes. The chloramphenicol-resistant colony was selected and cultured at 42°C for 6 h to eliminate the pKD46 plasmid. The obtained resistant colony was selected and cultured at 42°C for 6 h to induce the FLP recombination of the FRT sites and to eliminate the cat gene and the pCP20 plasmid. The final mutant was designated as E. coli DH5α Δnfi.

Nfi-F (upstream of the gene nfi) and Nfi-R (downstream of the gene nfi) primers were designed to indicate the change in the nfi gene. PCR was performed using Nfi-F and Nfi-R primers with the genomic DNAs of E. coli DH5α, E. coli DH5α cat Δnfi and E. coli DH5α Δnfi as templates. The PCR products were used in 0.8% agarose gel electrophoresis (100 V for 40 min) to determine their sizes.

SMRT sequencing of the PaP1 genome

The PaP1 genome was subjected to SMRT sequencing at the Institute of Medicinal Plant Development (Beijing, China) by using a PacBio RS DNA sequencer (Pacific Biosciences, Menlo Park, CA, USA; http://www.pacificbiosciences.com/) [27,28]. SMRT sequencing was performed in accordance with previously described protocols [6,14,15]. In brief, SMRTbell template libraries with DNA fragments of 2 kb were prepared [29,30]. Sequencing was then performed using one SMRT cell (http://www.pacificbiosciences.com/products/consumables/SMRT-cells/); zero-mode waveguide (ZMW) [31] signals were obtained. SMRT reads were mapped to the reference sequence of the PaP1 genome by using the BLASR software (https://github.com/PacificBiosciences/BLASR) [32] in accordance with standard mapping protocols. Interpulse durations (IPDs) were determined and processed as previously described [15,29,33] for all of the pulses aligned to each position in the PaP1 genome sequence. The modified bases were identified using SMRT Analysis Server v. 1.4.0 (Pacific Biosciences). The generated data sets are available for download at the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) [34] with the accession number of GSE50100 [GEO: GSE50100].

Bioinformatics analyses

DNASTar [35] was used to analyse the basic characteristics of the PaP1 genome sequence. The Internet tool tRNAscan-SE 1.21 [36] was used to predict tRNA genes in the DNA sequence with a cove score cutoff of 20. DNAMAN software (http://www.lynnon.com/) was used to analyse the localisation of the 20 contigs in the PaP1 genome and to graphically describe the result. The PanDa-Tox database (http://www.weizmann.ac.il/pandatox) [37] was used to analyse the putative DNA motifs that were toxic to bacteria in the PaP1 genome.

The raw modification calls of the PaP1 genomic DNA, produced using the SMRTPortal Analysis Platform v. 1.3.3 (Pacific Biosciences; details are available at http://www.pacbio.com/pdf/TN_Detecting_DNA_Base-Modifications.pdf), were collated as single Modifications.gff file. To predict modified motifs, we screened the Modifications.gff file by using publicly available R-scripts software (https://github.com/PacificBiosciences/motif-finding), as well as an online motif finding server (MEME, http://meme.nbcr.net/meme/cgi-bin/meme.cgi) [38]. PaP1 ORF48 was blasted against NCBI non-redundant protein sequences (nr) (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=Blast Search&BLAST_SPEC=&LINK_LOC=blasttab&LAST_PAGE=blast&LAST_PAGET=blast) to search probable correlations between ORF48 and methyltransferases. Protein sequences were subjected to multiple sequence alignments by using ClustalW [39] with default parameters and a phylogenetic tree was constructed and displayed using MEGA5 [40] with a neighbor-joining method [41].

Results

Shotgun strategy failed to obtain a complete PaP1 genome sequence

The PFGE result showed that the PaP1 genome is approximately 90 kb in length (Figure 1A). However, the sequencing results of the PaP1 genome by using the shotgun strategy only provided 20 contigs with various lengths (Figure 1B) and all of these 20 contigs could not be assembled in an integral genome. In addition, the overall length of these 20 contigs was approximately 47.7 kb, only almost half of 90 kb. We subjected the PaP1 genome to re-sequencing in another sequencing company by using the shotgun method. However, we obtained almost the same result, as in the first sequencing. We also performed primer walking [11] to fill the gaps, but we still failed to obtain the whole genome sequence of PaP1. Although we selected 216 clones of the random restriction library of the PaP1 genome for sequencing, all of the obtained sequences belong to the sequence sets of the 20 contigs.
PaP1 genome sequence obtained by Roche/454 sequencer

Using a Roche/454 DNA sequencer, we easily obtained the 91,715 bp whole genome sequence of PaP1. The PaP1 genome sequence and its annotations have been submitted to GenBank (Accession: HQ832595). On the basis of the comparative analysis results of the PaP1 genome sequence, we established a new genus named PaP1-like phages [9]. The PaP1 genome does not contain complicated secondary structures. To determine the relationship between the sequences obtained by the shotgun method and the Roche/454 DNA sequencer, we mapped the 20 contigs to the PaP1 genome sequence and found that all of the sequences of the 20 contigs are identical to the PaP1 genome sequence; however, gaps with various lengths are present among these contigs (Figure 2). The largest gap was approximately 10 kb, which was very large to be filled by primer walking [11]. The total sequence length of the 20 contigs was approximately 47.7 kb, only half of the whole PaP1 genome sequence (91.7 kb).

Single-molecule sequencing revealed modified bases in the PaP1 genome

The PaP1 genome could be successfully sequenced with the Roche/454 technique but not with the shotgun method. The shotgun method depends on the construction of a DNA library; by contrast, the Roche/454 technique is a non-library-dependent technique. Therefore, we hypothesised that the shotgun method failed possibly because E. coli DH5α, the host bacterium of the shotgun library construction, greatly repelled the inserted DNA fragments by endonucleases; the PaP1 genome may contain modified bases that may be the recognised targets degraded by endonucleases.

As such, we subjected the PaP1 genome to another sequencing by using a SMRT DNA sequencing technique [15] in 2013. In this procedure, the average sequence coverage of the SMRT sequencing reached 1,380-fold of the PaP1 genome (Additional file 1: Figure S1). We obtained IPD ratios of the 91,715 bases on both positive and reverse strands of the PaP1 genomic DNA. Among the IPD ratios, those of 7,557 bases (Additional file 2: Excel S1) exhibited typical signals of modified bases, including 51 of N-6-methyladenines (m6A), 152 of N-4-methylcytosines (m4C) and 7,354 other modified bases (unknown modified types because of the limitations of the current SMRT sequencing technique). Figure 3 shows the IPD ratios of both DNA strands in a section of the PaP1 genomic DNA by SMRT sequencing: A, B and C show the three typical instances (m6A, m4C and unknown modified bases)ograms.
base, respectively) of modified bases. Figure 4 shows an integral epigenetic map of the PaP1 genome, indicating the positions of m6As, m4Cs and unknown modified bases. These results indicated that the PaP1 genome contains numerous modified bases (7,557 in number), accounting for 8.2% of the total PaP1 genome sequence.

Methylome analysis of the PaP1 phage
We selected the top 10 modified motifs (with E-value ≤ 5.1e+004) from numerous motifs screened from the Modifications.gff file and analysed these motifs. We focused on motifs with the number of sites >10; hence, we only acquired three motifs (Figure 5). The consensus sequences of these three motifs are “5’-VAGRAGGH-3’,” “5’-AVASCMSRGC-3’,” and “5’-SMTSGKTARA-3’,” respectively. For these predicted motifs, only some of the sites found in the genome were detected as methylated; this result indicated that the methylated pattern and the methyltransferase(s) PaP1 used may be very complicated.

In silico analysis results revealed that the PaP1 ORF48 is a putative methyltransferase [9]. A total of 15 putative methyltransferases were found when the PaP1 ORF48 was compared with the protein database and the BlastP scores were ≥60 (Table 3). These 15 putative methyltransferases shared 22 identical amino acids (~21%) with the PaP1 ORF48 (Figure 6A). The phylogenetic tree further showed that the PaP1 ORF48 is closely related to the putative methyltransferase encoded by Pseudomonas phage JG004 and slightly related to methyltransferases encoded by bacteria (Figure 6B). However, we were unsure whether or not the PaP1 ORF48 is a putative methyltransferase because BlastP analysis results also suggested that the PaP1 ORF48 is related to phage portal proteins.

Digestion of the PaP1 genomic DNA by Endo V
Some enzymes of the host bacteria (E. coli DH5α) of the shotgun library construction probably target these modified bases because the PaP1 genomic DNA contains numerous modified bases. Hence, we doubted enzyme Endo V because this enzyme can recognise and degrade modified bases containing DNA molecules [42-45]. To confirm whether or not Endo V is responsible for the failure of the shotgun method, we used Endo V to digest the genomic DNA of PaP1. The results showed that the PaP1 genomic DNA formed a smear in the gel when this DNA was degraded with Endo V whereas the restriction endonuclease EcoRI cleaved the PaP1 genomic DNA into several independent fragments (Figure 7A). By contrast, the PaP3 genomic DNA [19], successfully sequenced using the shotgun method, cannot be degraded by Endo V under the same reaction condition (Figure 7B); this result suggested that no Endo V cutting site exists in the PaP3 genome.

![Figure 3 Trace of IPD ratio variations showing three instances of modified sequence regions in the PaP1 genome.](image-url)
Use of the $nfi^{-}$ mutant of E. coli DH5α as the host bacterium for shotgun library construction revealed the whole PaP1 genome sequence

To further validate the role of Endo V in the failure of the shotgun sequencing of the PaP1 genome and verify the aforementioned hypothesis, we knocked out the Endo V coding gene ($nfi$) of E. coli DH5α. The $nfi$ gene of E. coli DH5α genome was initially substituted with a donor DNA (containing chloramphenicol-resistant gene, $cat$) by using a $\lambda$-red recombination system; the $cat$ gene was then eliminated by FLP (a yeast-derived recombinase) recombination (Figure 8A). The PCR identification results showed that the sizes of the PCR products are correct (Figure 8B). These PCR products were sequenced and the results indicated that the $nfi$ gene was completely knocked out. This mutant was designated as E. coli DH5α Δ$nfi$ or the $nfi^{-}$ mutant of E. coli DH5α.

We used this mutant to construct the shotgun library of the PaP1 genomic DNA. The obtained shotgun reads were assembled into eight contigs that covered 92.3% of the PaP1 genome (Figure 8C) when the sequencing coverage reached 10-fold of the PaP1 genome. The length of the largest gap is <1.5 kb, which can be easily filled by primer walking [11]. Hence, the use of E. coli DH5α $nfi^{-}$ mutant as a host bacterium of shotgun library construction can overcome the inability of the shotgun method to complete the PaP1 genome sequence.

Discussion

In clone-based genome sequencing, some genomic DNA fragments cannot be cloned using E. coli; as a result, cloning gaps are retained when sequence reads are analysed. Although cloning-independent sequencing methods are available [5-7], the cause of the sequencing problem
remains unclear. Previous findings indicated that some restriction enzymes [46] and toxic small RNA are present in a shotgun-unclonable genome region. Furthermore, some DNA fragments in shotgun-unclonable regions suppress the growth of *E. coli* [37]. However, the PanDaTox database reveals that the PaP1 genome does not have any evident DNA motifs that are toxic to bacteria; in this study, a different viewpoint was proposed, in which the Endo V-mediated immunity of *E. coli* is responsible for the failure of the shotgun method to sequence a phage genome that contains modified bases.

This study was initiated when we found that the shotgun library method failed to sequence the genome of the PaP1 phage with a size of 90 kb in 2004. Several years later, Roche/454 sequencing method was established. We used the Roche/454 technique to sequence

![Figure 5](image-url) Modified motif prediction of the PaP1 genome. (A) Modification scores by cognate base. We selected the top 1,400 context sequences (with a score cutoff of 37) for analysis. A file of 1,400 sequences is suitable as the input for the online motif finding server (MEME). (B) Putative modified motifs determined from the PaP1 genome. A left “+” means forward and the corresponding right “-” means reverse complement.

| #  | Species                      | Subject ID               | Alignment length | Score | Identity (%) | E-value |
|----|------------------------------|--------------------------|------------------|-------|--------------|---------|
| 1  | Pseudomonas phage JG004      | gi|418488276| 156   | 317          | 99      | 2e-108  |
| 2  | *Haliangium ochraceum* DSM 14365 | gi|262194136| 137   | 66           | 32      | 3e-10   |
| 3  | *Lactococcus lactis*         | gi|480222890| 141   | 66           | 32      | 5e-10   |
| 4  | *Lactococcus lactis*         | gi|556501605| 128   | 66           | 34      | 6e-10   |
| 5  | *Lactococcus garvieae*       | gi|480228496| 130   | 66           | 32      | 6e-10   |
| 6  | *Lactococcus lactis* subsp. lactis* KLDS 4-0325 | gi|554465598| 141   | 65           | 32      | 8e-10   |
| 7  | *Haliangium ochraceum* DSM 14365 | gi|262193963| 137   | 64           | 32      | 3e-09   |
| 8  | *Enterococcus saccharolyticus* | gi|491614495| 128   | 62           | 31      | 9e-09   |
| 9  | *Myxococcus xanthus* DK 1622 | gi|108757550| 117   | 61           | 31      | 2e-08   |
| 10 | *Myxococcus xanthus*         | gi|521967607| 117   | 61           | 31      | 2e-08   |
| 11 | *Stigmatella aurantiaca* DW4/3-1 | gi|310819901| 117   | 61           | 32      | 3e-08   |
| 12 | *Myxococcus fulvus* HW-1     | gi|338536945| 117   | 61           | 32      | 3e-08   |
| 13 | *Myxococcus xanthus*         | gi|521967482| 117   | 61           | 32      | 3e-08   |
| 14 | *Myxococcus xanthus*         | gi|521967591| 117   | 61           | 31      | 4e-08   |
| 15 | *Stigmatella aurantiaca* DW4/3-1 | gi|310819798| 117   | 60           | 32      | 5e-08   |
the PaP1 genome again in 2008. We easily obtained the complete genome sequence (91,715 bp) of the PaP1 genome. As such, we wondered why the PaP1 genome could be successfully sequenced using Roche/454 technique but could not be sequenced using the shotgun method. In contrast to the Roche/454 strategy, the shotgun strategy requires shotgun library construction. Based on the principle difference of the two sequencing methods, our presumption was that *E. coli* DH5α, the host bacterium of the shotgun library construction, probably repel the inserted phage-DNA fragments via a particular immune mechanism.

The shotgun strategy has been successfully applied to sequence the genomes of many organisms, including bacteria, plants and animals, as well as viruses. The host bacteria of the constructed shotgun library construction, probably repel the inserted DNA fragments via a particular immune mechanism. The shotgun strategy has been successfully applied to sequence the genomes of many organisms, including bacteria, plants and animals, and animals. The host bacteria of the constructed shotgun library construction, probably repel the inserted DNA fragments via a particular immune mechanism. During evolution, bacteria most likely develop an immune mechanism that aims directly at these modified bases in exogenous DNA.

Several known bacterial immune mechanisms, such as R-M [55], T-A [56], Abi [57] and CRISPR-Cas [58] systems exist, but any of these mechanisms does not directly aim at varied modified bases in exogenous DNA. We then focused on the enzyme Endo V because this enzyme can recognise many kinds of modified bases in DNA strands [42,45,59]. The mechanism of Endo V activity is different from that of general restriction endonucleases in an R-M system because these restriction endonucleases of the R-M system generally recognise and cut at unmodified base sites [60]; by contrast, Endo V recognises and cuts at modified base sites. Endo V also exhibits endonuclease and exonuclease activities [61,62], which provide Endo V with a more effective DNA destruction activity than general restriction endonucleases.

Endo V was originally reported as a DNA repair enzyme [43,44,63] encoded by the *nfi* gene; most bacteria contain the *nfi* gene in their genome. This enzyme can...
recognise and cleave various modified bases and abnormal structures, such as deaminated bases, abasic (AP) sites, base mismatches, methylated bases, flap DNA, pseudo-Y structures and small insertions/deletions [42,45,59,63] in DNA molecules, with a cleavage site at the second phosphodiester bond in the 3′ direction from the recognition site; as a result, a nick with 5′-phosphate and 3′-hydroxyl groups is formed and DNA strands are greatly disrupted because of the exonuclease activity of this enzyme. To determine whether or not Endo V can destroy the PaP1 genomic DNA, Endo V (a product of E. coli nfi gene) was used to digest the PaP1 genomic DNA. The result indicated that Endo V degraded the PaP1 genomic DNA into a smear band (Figure 7A).

To further validate the role of Endo V in the failure of the shotgun sequencing of the PaP1 genome, we knocked out Endo V-coding nfi gene and constructed an nfi− mutant of E. coli DH5α. This mutant was then used as the host bacterium to construct the PaP1 genomic DNA shotgun library. Consequently, the obtained sequences covered 92.3% of the PaP1 genome when the sequencing amount of the PaP1 genome reached a 10-fold coverage and the largest gap between contigs was <1.5 kb (Figure 4), which
is very easy to close. This result further confirmed that the activity of Endo V is responsible for the failure of the shotgun sequencing of the PaP1 genome.

The SMRT DNA sequence of the PaP1 genome showed that 7,557 bases of this genome were substituted with modified bases, including 51 m6A, 152 m4C and 7,354 other modified bases (unidentified modified types, Figures 3 and 4). The positions of each modified base in the PaP1 genome (Figure 4) indicated the presence of modified bases in this genome. We also investigated the methylome of the PaP1 phage, which may be the first phage methylome revealed by SMRT technology; this methylome may be significant in future studies on phage biology and host interaction.

Conclusions
This work revealed the whole PaP1 genome sequence that contains numerous modified bases, provided complete information of the epigenetic information map of the PaP1 phage with 7,557 modified bases and investigated the methylome of PaP1. We found that the shotgun sequencing method is unsuitable for genomes containing many modified bases. To resolve this problem, we may use the nfi− mutant of E. coli DH5α as the host bacterium of DNA library construction. Moreover, we revealed a new mechanism of bacterial immunity to repel exogenous DNA by Endo V activity. Considering that bacteriophage is a virus infecting bacteria and modified bases are commonly found in a phage genome, the new mechanism of bacterial immunity we first demonstrated in this study, may be particularly necessary for bacteria to evade DNA invasion and retain their genetic stability.

Availability of supporting data
The nucleotide sequence of PaP1 phage was deposited in the GenBank database with the accession number of HQ832595 (http://www.ncbi.nlm.nih.gov/nuccore/HQ832595). The data sets supporting the results of this article are available in the NCBI GEO repository [34] with the accession number of GSE50100 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?&acc=GSE50100).

Additional files

Additional file 1: Figure S1. Depth of the SMRT sequencing coverage across the PaP1 genome. The window size is set at 200 bp. The average sequencing coverage reached approximately 1,380-fold of the PaP1 genome.

Additional file 2: Excel S1. IPD ratios of modified bases within the PaP1 genome, related to Figure 4. This Excel lists the details (including IPD ratios) of 7557 modified bases.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
The author(s) have made the following declarations about their contributions: FH, S. Lu and YT conceived and designed the experiments; S. Lu, S. Le, ML, CL, KZ, JH and HC performed the experiments; S. Lu, XR, JZ, L. Zou, QN, XJ, S. Li and JW analyzed the data; JW, XJ, QH, XZ, L. Zhang and GH contributed reagents/materials/analysis tools; FH and S. Lu wrote the paper. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by the National Natural Science Foundation of China (31070153) and the Chongqing Education Committee Foundation of China (101207). We would like to thank Professor Weiguo Cao (who works at the Department of Genetics and Biochemistry, South Carolina Experiment Station, Clemson University, Clemson, USA) for providing relevant information related to Endo V.

Received: 26 January 2014 Accepted: 16 September 2014 Published: 19 September 2014

References
1. Fiers W, Contreras R, Duehrick F, Haegeman G, Isertenant D, Merregaert J, Min Jou W, Molemans F, Raeymaekers A, Van den Berge A, Volckaert G, Ysebaert M: Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 1976, 260(5551):500–507.
2. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocumbe PM, Smith M: Nucleotide sequence of bacteriophage phi X174 DNA. Nature 1977, 268(5604):697–695.
3. Fleischmann RD, Adams MD, White O, Clayton RA, Kerlavage AR, Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM: Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995, 269(5223):496–512.
4. Fuchs TM, Brandt K, Starke M, Ratzi T: Shotgun sequencing of Yersinia enterocolitica strain W22703 (biotype 2, serotype O9): genomic evidence for oscillation between invertebrates and mammals. BMC Genomics 2011, 12:168.
5. Gupta PK: Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol 2008, 26(11):602–611.
6. McCarthy A: Third generation DNA sequencing: pacific biosciences’ single molecule real time technology. Chem Biol 2010, 17(11):767–76.
7. Shendure J, Ji H: Next-generation DNA sequencing. Nat Biotechnol 2008, 26(10):1135–1145.
8. Krebes J, Morgan RD, Bunk B, Spoer C, Luong K, Parulel R, Anton BP, Konig C, Josienschans C, Overmarn J, Roberts RJ, Korlach J, Suerbaum S: The complex methylome of the human gastric pathogen Helicobacter pylori. Nucleic Acids Res 2013, 42(4):2415–2432.
9. Lu S, Le S, Tan Y, Zhu J, Li M, Rao X, Zou L, Li S, Wang J, Jin X, Huang G, Zhang L, Zhao X, Hu F: Genomic and proteomic analyses of the terminally redundant genome of the pseudomonas aeruginosa phage PA01: establishment of genus PA01-like phages. PLoS One 2013, 8(3):e62933.
10. Le S, He XS, Tan YL, Huang GT, Zhang L, Lux R, Shi WY, Hu FQ: Mapping the tail fiber as the receptor binding protein responsible for differential host specificity of pseudomonas aeruginosa bacteriophages PaP1 and JG004. PLoS One 2013, 8(7):e68562.
11. Benes V, Kligler C, Voss H, Paabo S, Anborgo W: Direct primer walking on P1 plasmid DNA. Biotechniques 1997, 23(1):98–100.
12. Zheng Z, Advani A, Meleifors O, Glavas S, Nordstrom H, Ye W, Enstrand L, Anderson AF: Titration-free 454 sequencing using Y adaptors. Nat Protoc 2011, 6(9):1367–1376.
13. Clark MS, Thorne MA, Vieira FA, Cardoso JC, Power DM, Peck LS: Insights into shell deposition in the Antarctic bivalve Laternula elliptica: gene discovery in the mantle transcriptome using 454 pyrosequencing. BMC Genomics 2010, 11:362.
14. Davis BM, Chao MC, Waldor MK: Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. Curr Opin Microbiol 2013, 16(2):192–198.

15. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olives EC, Clark TA, Korlach J, Turner SW: Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 2010, 7(5):461–465.

16. Murray TA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, Fomenkov A, Turner SW, Korlach J, Roberts RJ: The methylomes of six bacteria. Nucleic Acids Res 2012, 40(22):11450–11462.

17. Sambrook J, Russell DW: The Condensed Protocols from Molecular Cloning: A Laboratory Manual. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press; 2006.

18. Sun WZ, Tan XL, Ja M, Hu XM, Yao XC, Hu FQ: Functional characterization of the endolysin gene encoded by Pseudomonas aeruginosa bacteriophage PaP1. Afr J Microbiol Res 2010, 4(10):933–939.

19. Tan Y, Zhang K, Rao X, Jin X, Huang J, Zhu J, Chen Z, Hu X, Shen X, Wang L, Hu F: Whole genome sequencing of a novel temperate bacteriophage of P. aeruginosa: evidence of tRNA gene mediating integration of the phage genome into the host bacterial chromosome. Cell Microbiol 2007, 9(2):479–491.

20. Govind R, Fralick JA, Rolfe RD: Genomic organization and molecular characterization of Clostridium difficile bacteriophage PhiCD119. J Bacteriol 2006, 188(7):2568–2577.

21. Casas V, Rohwer F: Phage metagenomics. Methods Enzymol 2007, 432:259–268.

22. Vandersteegen K, Kropinski AM, Nash JH, Noben JP, Hermans K, Lavigne R, Roberts RJ, Korlach J, Roberts RJ: Using single-molecule, real-time DNA sequencing. Nat Rev Microbiol 2012, 10(6):401–408.

23. de la Bastide M, McCombie WR: Short read fragment assembly of bacterial functional genomics data sets. Nucleic Acids Res 2008, 36(13):3983–3991.

24. Fierer N, Robertson CL, Serova N, Davis S, Soboleva A: NCBJ GEO: archive for functional genomics data sets-update. Nucleic Acids Res 2013, 41(Database issue):D1991–D1995.

25. Rosset I, Scheuch M, Hober D, De Regge N, Caij AB, Vandenbussche F, Van Borm S: Dblase SBPA-next generation sequencing confirms Schmallenberg virus in Belgian field samples and identifies genetic variation in Europe. PLoS One 2012, 7(7):e41967.

26. Datsenko KA, Wanner BL: Gene disruption in Escherichia coli: TcR

27. Cherepanov PP, Wackernagel W: The Condensed Protocols from Molecular Cloning: A Laboratory Manual. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press; 2006.

28. Datsenko KA, Wanner BL: Gene disruption in Escherichia coli: TcR

29. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003, 31(13):3997–3950.

30. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGAS: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 2011, 28(4):1595–1609.

31. Gates FT 3rd, Linn S: Endonuclease V with DNA reveal initiation of deaminated adenine repair. Nat Struct Mol Biol 2009, 16(6):138–143.

32. Datsenko KA, Wanner BL: Gene disruption in Escherichia coli: TcR

33. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGAS: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 2011, 28(4):1595–1609.

34. Gates FT 3rd, Linn S: Endonuclease V with DNA reveal initiation of deaminated adenine repair. Nat Struct Mol Biol 2009, 16(6):138–143.

35. Datsenko KA, Wanner BL: Gene disruption in Escherichia coli: TcR

36. Schattner P, Brooks AN, Lowe TM: The trnAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 2005, 33(Web Server issue):W686–W689.

37. Kimmelman A, Levy A, Sberro H, Kidron S, Leavitt A, Amitai G, Yoder-Himes DR, Wurtzel O, Zhu YW, Rubin EM, Sorok R: A vast collection of microbial genes that are toxic to bacteria. Genome Res 2012, 22(4):802–809.

38. Bailey TL, Boden M, Busa FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 2009, 37(Web Server issue):W202–W208.

39. Chenra R, Sugawara H, Koihe T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003, 31(13):3997–3950.

40. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGAS: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 2011, 28(4):1595–1609.

41. Gates FT 3rd, Linn S: Endonuclease V with DNA reveal initiation of deaminated adenine repair. Nat Struct Mol Biol 2009, 16(6):138–143.

42. Datsenko KA, Wanner BL: Gene disruption in Escherichia coli: TcR

43. Gates FT 3rd, Linn S: Endonuclease V with DNA reveal initiation of deaminated adenine repair. Nat Struct Mol Biol 2009, 16(6):138–143.

44. Datsenko KA, Wanner BL: Gene disruption in Escherichia coli: TcR

45. Gates FT 3rd, Linn S: Endonuclease V with DNA reveal initiation of deaminated adenine repair. Nat Struct Mol Biol 2009, 16(6):138–143.

46. Schattner P, Brooks AN, Lowe TM: The trnAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 2005, 33(Web Server issue):W686–W689.

47. Kimmelman A, Levy A, Sberro H, Kidron S, Leavitt A, Amitai G, Yoder-Himes DR, Wurtzel O, Zhu YW, Rubin EM, Sorok R: A vast collection of microbial genes that are toxic to bacteria. Genome Res 2012, 22(4):802–809.

48. Bailey TL, Boden M, Busa FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 2009, 37(Web Server issue):W202–W208.

49. Chenra R, Sugawara H, Koihe T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003, 31(13):3997–3950.
59. Childs JD, Paterson MC, Smith BP, Gentner NE: Evidence for a near UV-induced photoproduct of 5-hydroxymethylcytosine in bacteriophage T4 that can be recognized by endonuclease V. Mol Gen Genet 1978, 167(1):105–112.

60. Bickle TA, Kruger DH: Biology of DNA restriction. Microbiol Rev 1993, 57(2):434–450.

61. Huang J, Lu J, Barany F, Cao W: Multiple cleavage activities of endonuclease V from Thermotoga maritima: recognition and strand nicking mechanism. Biochemistry 2001, 40(30):8738–8748.

62. Majorek KA, Bujnicki JM: Modeling of Escherichia coli Endonuclease V structure in complex with DNA. J Mol Model 2009, 15(2):173–182.

63. Rosnes I, Rowe AD, Vik ES, Forstnæ RJ, Alseth I, Bjoras M, Dalhus B: Structural basis of DNA loop recognition by endonuclease V. Structure 2013, 21(2):257–265.