Research Article

SeqWord Motif Mapper: A Tool for Rapid Statistical Analysis and Visualization of Epigenetic Modifications in Bacterial Genomes

Christophe M. J. Lefebvre, Rian E. Pierneef, and Oleg N. Reva *

Centre for Bioinformatics and Computational Biology, *Dep. of Biochemistry, Genetics and Microbiology, University of Pretoria, South Africa*

Correspondence to Oleg N. Reva: *oleg.reva@up.ac.za (O.N. Reva)*
https://doi.org/10.1016/j.jmb.2025.169307
Edited by Rita Casadio

Abstract

Genomic methylation in bacteria plays a crucial role in gene regulation, chromosome replication, pathogenicity, and defense against phages. While single-molecule real-time (SMRT) sequencing technologies have advanced the detection of epigenetically modified bases, the statistical analysis of their distribution and the possible roles they play in bacterial cells remains challenging. To address this gap, we developed SeqWord Motif Mapper (SWMM), a computational tool designed for the statistical analysis and visualization of bacterial methylation patterns. SWMM utilizes PacBio sequencing data to identify sequence coverage, methylation motif distribution, and putative functional associations. Implemented in Python 3.9, the tool is platform-independent and requires minimal dependencies, making it accessible to a wide range of users. The SWMM command-line interface and a web-based version of the program facilitate the exploration of epigenetic modifications across bacterial genomes. Through case studies on different bacterial and archaeal taxa, we demonstrated that genome methylation in microorganisms extends beyond canonical sites and possibly influences gene expression, adaptation, and genome architecture. The tool enables detailed statistical evaluation of methylation motif distribution and provides insights into the potential regulatory roles of epigenetic modifications in bacterial genomes. SWMM is freely available at https://begp.bi.up.ac.za, with source code hosted on GitHub at https://github.com/chrilef/BactEpiGenPro.

Introduction

Epigenetics is a rapidly developing field of genomic studies, revealing new facets of genetic inheritance and gene expression regulation mechanisms that are not associated with genetic mutations or transcription factors. Significant advancements in epigenetic studies have been made in recent years due to third-generation sequencing technologies, such as PacBio Single-Molecule Real-Time Sequencing (SMRT) and Oxford Nanopore Sequencing Technologies (ONT), which enable the detection of chemically modified nucleotides alongside genome sequencing [1]. The most common type of epigenetic modification is methylation of nucleotides. However, epigenetic modifications also include other chemical reactions, such as the oxidation of nucleotides in genomic DNA [2].

Methylation is a key epigenetic mechanism driving several important processes in bacteria, such as defense against phages, chromosome replication, mismatch repair, gene regulation, and pathogenicity [3–6]. Fundamentally, DNA methylation is mediated by a diverse group of enzymes known as methyltransferases (MTases), which catalyze the transfer of methyl groups from S-adenosyl-L-methionine (SAM) to a positional

carbon atom, resulting in one of three methylated bases: N6-methyladenine (m6A), N4-methylcytosine (m4C), and 5-methylcytosine (5mC), although the latter is less frequently observed in prokaryotes and archaea [1,7]. MTases typically methylate specific nucleotide sequences termed canonical motifs; however, other studies showed that canonical motifs are semi-conserved [8]. Restriction-modification (RM) systems often are acquired horizontally forming strain-specific patterns of methylated motifs [9–11]. This behavior is further influenced by interactions with DNA-binding proteins. For example, Peterson and Reich (2008) detailed a phase variation model in uropathogenic *E. coli*, where on/off switching of *pap* operon expression was determined by the competitive binding of G*a*TC sites between the leucine-responsive regulatory protein (LRP) and DNA-adenine-specific Dam MTase. It was hypothesized that this dynamic interaction is dependent on nutrient availability, which improves survivability in nutrient-limited scenarios [12]. A similar observation in phylogenetically distinct *Bacillus* species implicates methylation as a mechanism that confers clonal diversity and enhances adaptability in different habitats by altering global gene expression [13,14]. Understanding the underlying roles of methylation is therefore crucial, as it may have significant applications in areas such as infectious disease research and agriculture.

Classical approaches for methylation profiling include laboratory-based methods like restriction analysis and bisulphite sequencing [15,16]. However, over 6,000 sequenced bacterial genomes are known to encode MTases [1], necessitating the development of more rapid and versatile tools. With the advent of third-generation sequencing technologies such as PacBio Single-Molecule Real-Time Sequencing (SMRT) and Oxford Nanopore Sequencing Technologies (ONT), epigenetic signals can be easily detected, and researchers have leveraged these platforms to successfully infer methylation motifs without prior knowledge of their sequence contexts [17,18]. The PacBio SMRT technology has become the gold standard for detecting epigenetically modified nucleotides, fully reproducing genome methylation patterns predicted by traditional chemical methods for identifying modified nucleotides in bacteria and archaea [19,20]. However, several computational tools have been developed recently for detecting epigenetic modifications using ONT-generated DNA sequencing data [21,22].

As previously mentioned, the biological complexity of methylation can constrain mapping genomic sites, as strain-specific methylation may lead to non-canonical sites with alternative biological roles. Software such as ipdSummary and MotifMaker, provided by the SMRT Link package (https://www.pacb.com/support/software-downloads/), can be utilized to discover epigenetically modified nucleotides and predict canonical motifs, which can then be compared against curated databases like REBASE [23]. However, there is a need for additional software tools to perform downstream analysis and visualization of the output files generated by SMRT Link programs. Here, we introduce SeqWord Motif Mapper, a tool for the statistical analysis and visualization of bacterial methylation patterns. A recently published software tool BacMethy [24], is, to our knowledge, the only available resource that addresses analysis of methylomics patterns in bacterial genomes, although it lacks functionality to analyze the distribution of minor or combinatorial methylation motifs, which may be overlooked by the MotifMaker program.

Methods

Software implementation and dependencies

The SWMM program is implemented in Python 3.9 and is compatible with all common operating systems, requiring only minimal dependencies for its operation. These dependencies include the standard Python 3 library *NumPy* (v. 1.26.4 or later), while statistical analyses are implemented using the standard *SciPy* library (v. 1.11.2 or later). The program saves output graphs either in SVG format or as HTML files with embedded SVG code. Converting output files to other graphical formats (e.g., EPS, PDF, and various raster formats) requires the installation of Poppler, Ghostscript and GTK software packages, as well as additional Python libraries: *Pango*, *GDK-PixBuf*, *CairoSVG*, and *pdf2image*. If the necessary software packages for generating the requested graphical output format are not installed, the program defaults to HTML output. To simplify dependency installation, a conda minimal environment file *environment_minimal.yml* was created and made available for downloading.

General statistical approaches

To identify potential associations between the locations of modified nucleosides within genomic DNA and the general properties of DNA loci, the program uses a sliding window approach to calculate local GC-content and GC-skew (Eqs. (1) and (2).

$$GC\_content = \frac{\sum G + \sum C}{SWL} \quad (1)$$

$$GC\_skew = \frac{\sum G - \sum C}{\sum G + \sum C} \quad (2)$$

where $\sum$G and $\sum$C represent the total counts of guanine and cytosine residues in a sliding window and SWL refers to the sliding window length.

Additionally, the program can identify the locations of horizontally acquired genetic islands

by incorporating applications from the previously designed SeqWord Gene Island Sniffer program [25]. For the statistical evaluation of the biased distribution of modified bases across genomic islands and the host genome, or between non-coding, coding and promoter regions, Z-values representing the deviation of the observed numbers of modified bases within regions from expected numbers were calculated and normalized by Eq. (3), assuming Poisson-distributed counts:

$$Z = \frac{count_{observed} - \mu}{\sqrt{\mu}} \quad (3)$$

where μ represents the expected count of modified bases, calculated as the ratio of the length of the regions of interest to the total length of the genome. The statistical significance of Z-deviations was confirmed by estimation *p*-values using the survival function (SF) of the standard normal distribution at |Z|, as implemented in the *stats.norm.sf* function of the *SciPy* library.

Spearman correlation between modified base counts within sliding windows and GC-content or GC-skew was calculated using the *stats.spearman* function of the *SciPy* library. The statistical significance of the calculated correlation was tested by estimating the confidence interval with an alpha of 0.05, using a bootstrap approach with 1,000 replicates. Confidence interval boundaries were calculated using the *percentile* function of the *NumPy* library. The correlation coefficient was considered significant if both the lower and upper confidence interval boundaries were either above or below zero.

When a genome consists of multiple replicons or contigs, the *p*-value of the deviation of observed counts of modified bases per contig from the expected counts, based on the contig lengths, was calculated using the *stats.chi2_contingency* function of the *SciPy* library.

Data availability

The SWMM source code is publicly available on GitHub at https://github.com/chrilef/BactEpiGenPro under the GNU General Public License (GPL). Included in the files is a pre-configured Conda (https://anaconda.org/) environment which may be imported to effortlessly manage all dependencies. The program can also be run online at https://begp.bi.up.ac.za. Beta-versions of the program were previously used to visualize methylation patterns in several projects [8,10,13,26–29].

Results

Preparatory steps

The program SeqWord Motif Mapper (SWMM) was designed to visualize patterns of epigenetically modified bases in genome-scale bacterial sequences by mapping long DNA reads generated by SMRT PacBio, while retaining base call kinetics data. Base call kinetics data is recorded during SMRT sequencing and stored in the resulting BAM file under the tags 'ip' (inter-pulse duration) and 'wp' (width of pulse). It should be noted that FASTQ files generated from the initial BAM files do not include kinetic data records. Additionally, the retention of kinetic data in BAM files is optional and can be disabled by the sequencing provider. It is therefore imperative to ensure that the aims of sequencing are properly discussed with the service provider.

Due to rapid advancements in PacBio sequencing technology and data processing, the format of base call kinetics data is evolving and varies between different platforms, such as Sequel and Revio. Ensure that the latest version of the SMRT Link package is used. Starting from SMRT Link v.13.0, the program Primrose is used instead of ipdSummary for calling epigenetically modified bases. Primrose produces GFF output files compatible with the program SWMM.

An example of a qsub script for calling methylated sites and canonical methylation motifs is shown below:

```
# Load the latest version of smrtlink
module load smrtlink_12.0.0.177059
# Create XML dataset based on one or multiple raw BAM files
dataset create --type SubreadSet/path/ subreadset.xml /path/1.bam2.bam 3.bam
# Create MMI index of the reference genome in FASTA format
pbmm2 index /path/reference.fa/path/reference.mmi
# Align and sort raw PacBio reads against the reference
pbmm2 align --sort /path/reference.mmi/path/subreadset.xml /path/alignment.bam
# Identify modified nucleotides and store the report in a GFF file
ipdSummary/path/alignment.bam --reference /path/reference.fa --identify m6A,m4C --gff/path/dnamod.gff
# Search for canonical motifs
motifMaker find --fasta/path/reference.fa --gff/path/dnamod.gff --minScore 20 --output /path/motifs.csv
```

The GFF file created by the ipdSummary program, containing the locations of modified nucleotides along with statistical validation data, is used as an input file for the SWMM program. Another required input file is the reference genome sequence. To utilize the full functionality of the program, the reference genome should be annotated and provided in GenBank (GBK) format. The CSV and SVG output files, generated by SWMM, statistically evaluate the distribution of epigenetically modified nucleosides and associated canonical motifs.

Program overview

The core functions of SWMM include several steps such as input parsing, parsing GFF input file for modified base prediction, parsing input GBK file for sequence and gene annotation data, verification of modified base locations, statistical validation and generation of requested graphical output files, as it is shown in the program workflow in Figure 1.

It should be highlighted that the locations of modified bases in the GFF files produced by the program ipdSummary do not exactly match to corresponding loci in the reference genome file. These mismatches between the original reference sequence and the resulting GFF file with locations of modified nucleotides is likely unexpected by the users as this problem is not documented in the programs used in the base modification calling pipelines. The problem can stem from improper alignment of reads against the reference sequence by the *pbmm2* program based on the *minimap2* algorithm, which uses split or chimeric alignments to handle large insertions/deletions and structural variation [30]. Moreover, IPD signals are sensitive to alignment errors. If the reads are not aligned perfectly, the reported modification sites may not correspond accurately to positions in the reference genome. Bases reported in the GFF outputs as modified may at times not be found in the reference sequence.

To address these discrepancies, SWMM uses BLASTN alignment of consensus sequences provided in ipdSummary generated GFF files. These consensus sequences are 41 bp in length with modified bases in their centers. The program SWMM searches for a consensus sequence match near the region of the predicted modified base allowing a specified number of mismatches in the argument *--blast_context_mismatch* including or excluding mismatches within the canonical motif sequence by setting the argument *--blast_motif_mismatch*. The program's runtime averages a few minutes but may vary depending on the size of the genome, the complexity of the motif, and the number of modified bases. The verification of modified base locations may be time consuming and depends on the number of verifications. The argument *--maximum_sites* specifies the limit of locations to verify (10,000 by default), but it can be set to zero to allow an unrestricted verification, which may utilize additional computational time and power.

SWMM generates three types of graphical outputs: (1) circular methylation map (MM) graph visualization of the distribution of DNA motifs and individual methylated bases across bacterial replicons (chromosomes and plasmids); (2) dotplot (DP) visualization of base modification scores and local sequence depth (coverage) values as predicted by the program ipdSummary; and (3) statistical panel (SP) providing the results of validation to assess the non-random distribution of epigenetically modified nucleosides across different genomic regions and replicons. SP graphs are combined with MM or DP graphs when requested together but may also be generated separately upon user request.

Running SeqWord motif mapper

SWMM is available as a standalone program for local usage or can be accessed online at https://begp.bi.up.ac.za. The local version of the program is stored to a folder on the computer, which contains the Conda environment file *environment_minimal.yml*, the Python executable file *run.py*, and three subfolders: *lib* with necessary Python module files, *input* with user input files, and *output* where the resulting output files will be stored. Project subfolders can be created in the folder *input*. The project subfolder name must be indicated using the argument *--project_directory*. If the program is instructed to read input files from the project directory, the program will create a subfolder with the project name in the folder *output*, where the resulting output files will be stored.

In addition to the GFF and GBK input files, the program requires setting several arguments, such as methylation motifs, specific nucleotides or the type of base methylation to identify. Additional values may be supplied by the user to adjust the strictness of the search. These parameters can be configured by the user either through the command-line interface, via a parameter-setting menu displayed in the command-prompt window (Figure 2), or by utilizing the Web-based user interface (https://begp.bi.up.ac.za). Several important program settings are discussed in more detail with example usage in subsequent sections. A detailed description of all program parameters can be found at https://github.com/chrilef/BactEpiGenPro.

Practical examples of SWMM application

The program was constructed with the aim of gaining insight into the possible biological roles of epigenetically modified bases in the genomic DNA of bacteria and archaea. Several examples of the program's practical applications are shown below. Listings of command prompts used to generate the example output files are shown in Suppl. Table S1.

Introducing canonical and non-canonical patterns of epigenetic modifications in bacterial genomes. Methylation of the microbial genome is associated with the activity of specific enzymes, methyltransferases (MTases), which are often part of restriction-modification (RM) systems that provide bacteria with defense mechanisms against phages and parasitic mobile genetic elements.

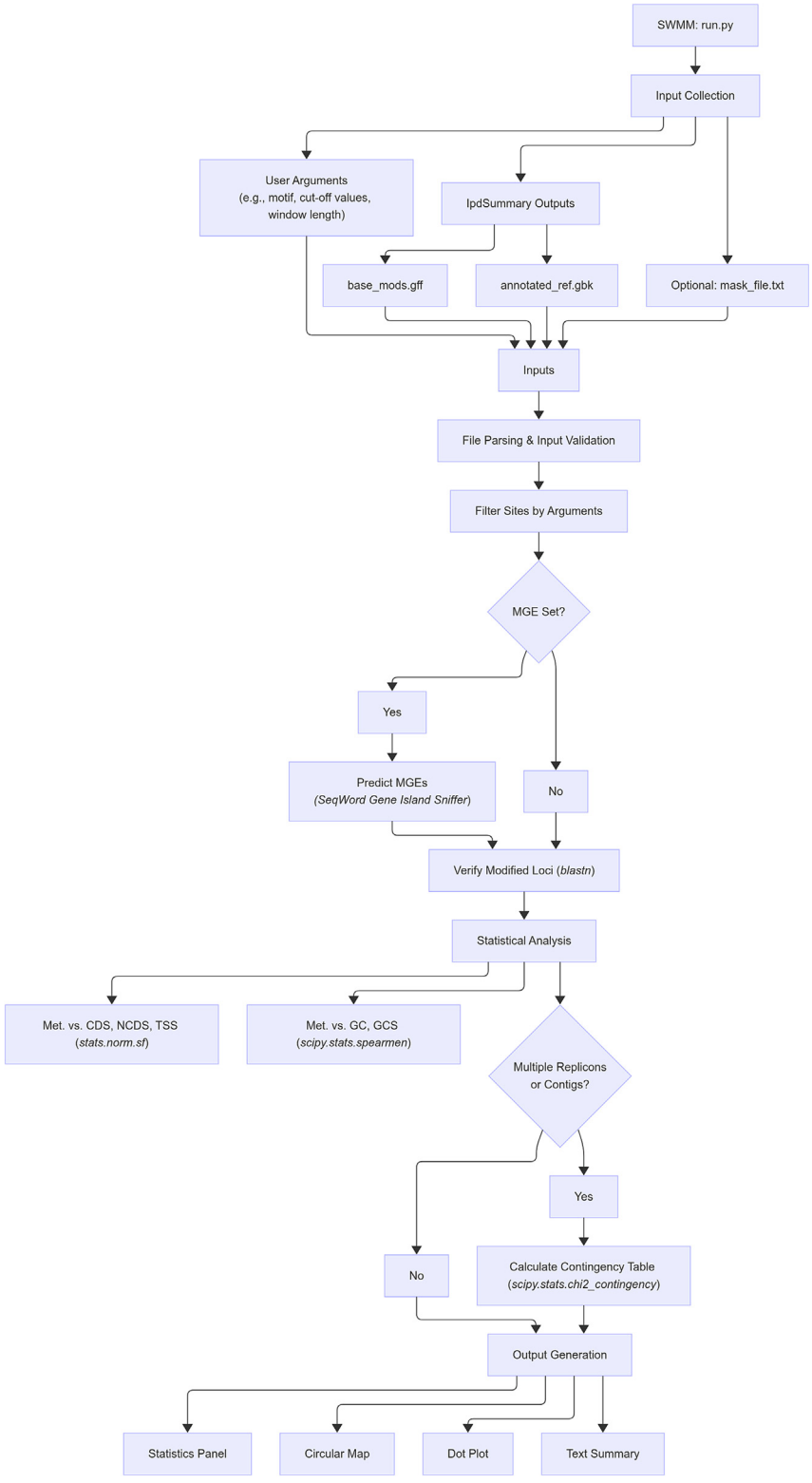

Figure 1. Workflow of the SeqWord Motif Mapper program.

However, methylation of genomic DNA is often carried out by orphan MTases, which are either remnants of former RM systems, or exclusively exist as solitary MTases. One well-known solitary MTase is Dam, which is abundant among Gram-negative enterobacteria but can also be found in many other microorganisms [10]. MTases recognize specific short DNA sequences known as

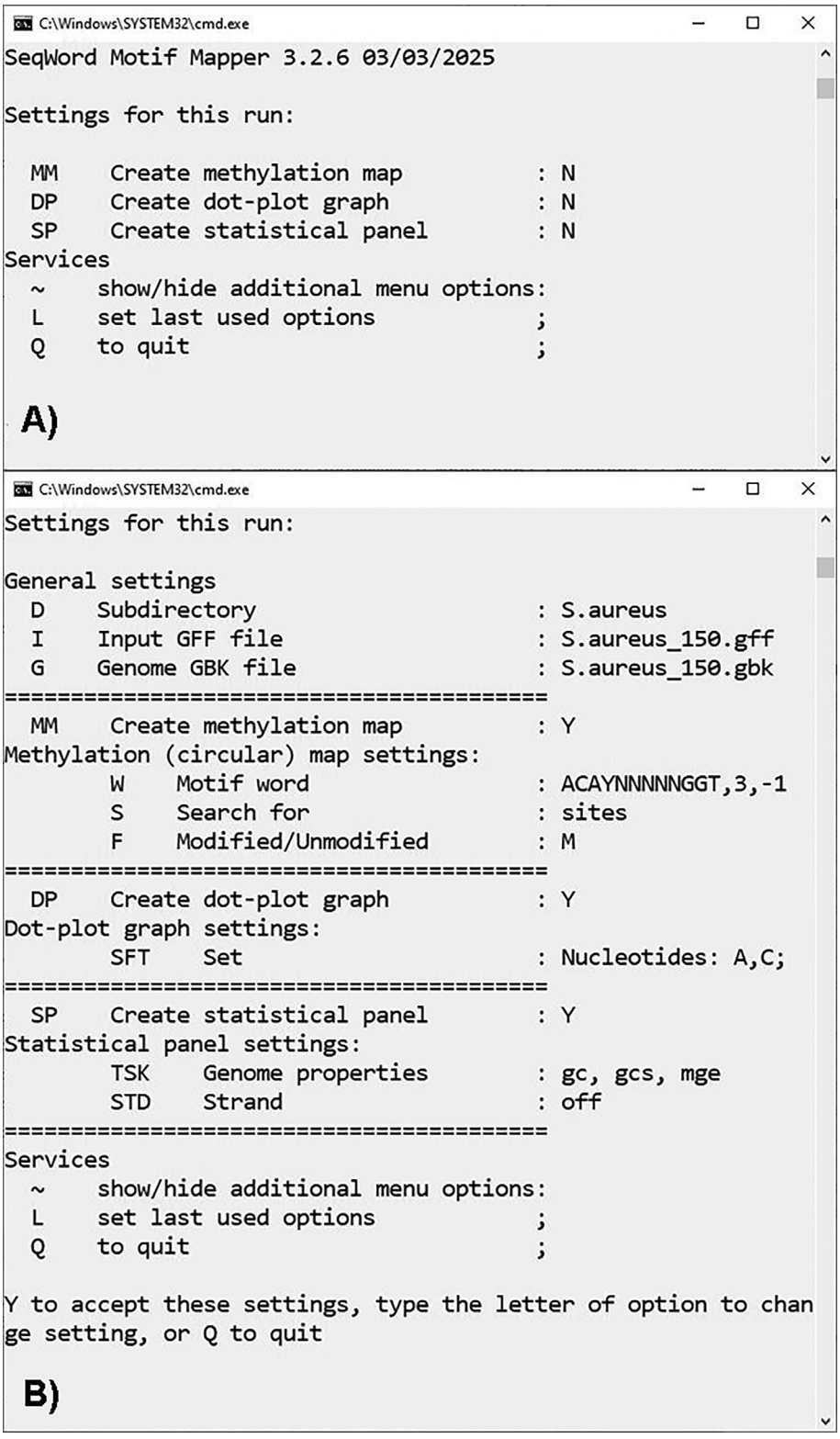

Figure 2. Command prompt interface of the SeqWord Motif Mapper program. (A) Upon start up, the program prompts the user to specify the types of graphs to be generated. The command L, followed by Enter <L + Enter>, can be used to set arguments from the previous successful program run. (B) Once the output graph types are selected, the program displays common arguments applicable to the selected outputs. To display all available arguments, press <~+Enter>.

canonical motifs. They bind to canonical motifs and methylate specific adenine or cytosine residues within the motif on one or both DNA strands. Dam MTase performs methylation at G*at*C palindromic motifs, which is widely considered as an epigenetic mechanism of gene expression regulation [5,6]. As such, the methylated nucleotides, cytosine and adenine, and the respective guanine and thymidine residues opposite the methylated nucleotides on the reverse-complement DNA strand, are depicted in the canonical motifs by lowercase cursive letters.

The abundance and distribution of methylated G*at*C motifs in *E. coli* ATCC BAA-39 [CP042865] were analyzed using the SWMM program, as shown in Figure 3A. The program was executed using the command shown in Listing 1 (Suppl. Table S1).

In total, 19,210 methylated adenine residues were found associated with the G*at*C motifs, predominantly on both DNA strands. The methylated motifs were abundant in protein-coding sequences but rare in the 120 bp TSC-upstream regions and 13 identified genetic islands. These deviations from the expected numbers of methylated motifs were statistically significant. The preferential occurrence of methylated adenine residues within GC-rich protein-coding regions is also reflected in a positive correlation with GC content. These motifs were randomly distributed across the chromosome, as evidenced by the absence of correlation with GC skew.

The command prompt shown in Listing 1 can be modified to display epigenetically modified bases that are not associated with the canonical motifs. To perform this task, the *--dotplot_motifs* argument should be set using the canonical motif with a negative sign: *−GatC,2,-2*. Additionally, the *--nucleotides* argument was set to *A,C* to display only adenine and cytosine modified bases. Figure 3B shows the resulting pattern of non-canonical epigenetically modified bases. To some extent, the non-canonical modification of nucleotides can be explained by inaccurate recognition of canonical motifs by MTases. In total, 6,828 cytosine residues were predicted to be m4C-methylated. The nature of other modifications remains unclear.

Despite the supposedly random nature of non-canonical modifications, their distribution was biased. The frequency of non-canonical modifications was significantly reduced in coding regions and the 120 bp TSC-upstream regions. However, they were abundant in non-coding regions, predominantly on the leading replichore, as indicated by a positive correlation with GC skew.

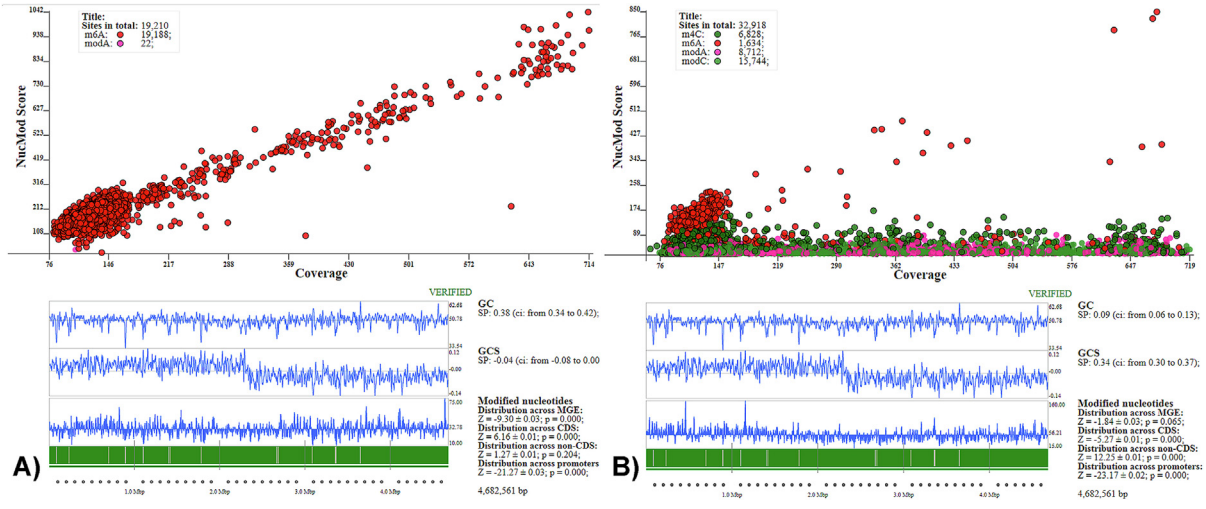

Figure 3. Dot-plot representation of (A) methylated adenine residues associated with the canonical G*at*C motif and (B) non-canonical epigenetically modified adenine and cytosine residues in the genome of *E. coli* ATCC BAA-39. Each node in the plots represents a modified base, characterized by NucMod score and local coverage values, as estimated by the ipdSummary program. The statistical panels below the dot-plot graphs illustrate the distribution of various canonical motifs across the *E. coli* ATCC BAA-39 chromosome. The graphs within the statistical panels, from top to bottom, depict GC content, GC skew, and the number of methylated bases within 8 kbp sliding windows, advancing along chromosomal sequences in 2 kbp steps. In the lower histograms, modified base densities are represented by bars extending above and below the average density line. Chromosomal coordinates are displayed at the bottom of the panels, with identified mobile genetic element (MGE) insertions indicated by pink bars. The results of the statistical analysis are presented on the right side of the panels and include Spearman correlation (SP) values with confidence intervals (ci) between GC content, GC skew, and the number of modified bases within sliding windows. Additionally, Z-scores, standard errors, and estimated *p*-values are provided for the biased distribution of modified bases across MGEs, core genomic regions, and between coding, non-coding, and 120 bp TSC-upstream regions.

Analysis of methylation coverage and identification of unmethylated bases within canonical motifs. The coverage of Dam methylation through the available G*at*C motifs in the genome of *E. coli* ATCC BAA-39 can be visualized using the command prompt show in Listing 2 (Suppl. Table S1).

The output of Listing 2 is shown in Figure 4A. The G*at*C methylation appeared to be highly effective: 38,423 sites were methylated out of 38,442, leaving only 19 unmethylated bases. However, the presence of conditionally modified bases is considered important in terms of their involvement in gene regulation [6]. To visualize the distribution of unmethylated GATC motifs shown in Figure 4B, the argument *--modified_or_unmodified* in Listing 2 must be set to *U* (for unmethylated).

In eight GATC motifs, adenine residues remained unmethylated on both DNA strands, whereas in three other motifs, methylation occurred only on one strand: G*a*TC. To display completely methylated or completely unmethylated motifs, the argument *--sites_or_motifs* in Listing 2 must be set to *motifs*.

Analysis of complexity of Dam methylation using the circular map graph. The identification of m4C-methylated cytosine residues in the genome of *E. coli* ATCC BAA-39, where no other MTases were found except for the Dam MTase, can be explained by the previously unknown complexity of Dam methylation, which involves cytosine methylation in non-palindromic *c*RGKG*at*C motifs [27].

The program SWMM allows for the visualization of complex motifs comprising ambiguous base settings and multiple methylation sites. The result of setting the argument *--cmap_motif* in Listing 2 to *cRGKGatC,1,6,-2* is shown in Figure 5A. Out of 2,235 available sites, 1,749 remained unmethylated amounting to 486 cytosine residues. The distribution of methylated *c*RGKG*at*C motifs was similar to that of G*at*C motifs (Figure 3A). Analysis of the distribution of *c*RGKG*at*C motifs revealed that they formed even more complex super-palindromic *c*RGKG*at*CMCY*g*,1,6,-1,-6 motifs, as shown in Figure 5B.

To determine whether cytosine methylation in *c*RGKG*at*C and *c*RGKG*at*CMCY*g* motifs is a strain-specific or common property, other bacterial strains containing Dam MTases were analyzed [10]. These strains include *E. coli* 3/145 [CP082827], *E. coli* 19/278 [CP082830], *Klebsiella pneumoniae* 13/97 [CP082805], *K. pneumoniae* 20/245 [CP082796] and *Streptococcus pneumoniae* PHRX1 [CP082820]. The results for the *c*RGKG*at*C and *c*RGKG*at*CMCYg motifs are shown in Supplementary Figures S1 and S2, respectively. The results demonstrate that cytosine methylation is common in microorganisms with Dam MTase, particularly in *K. pneumoniae* genomes. However, in the Gram-positive strain *S. pneumoniae* PHRX1, this methylation was less frequent compared to enterobacteria. Another com-

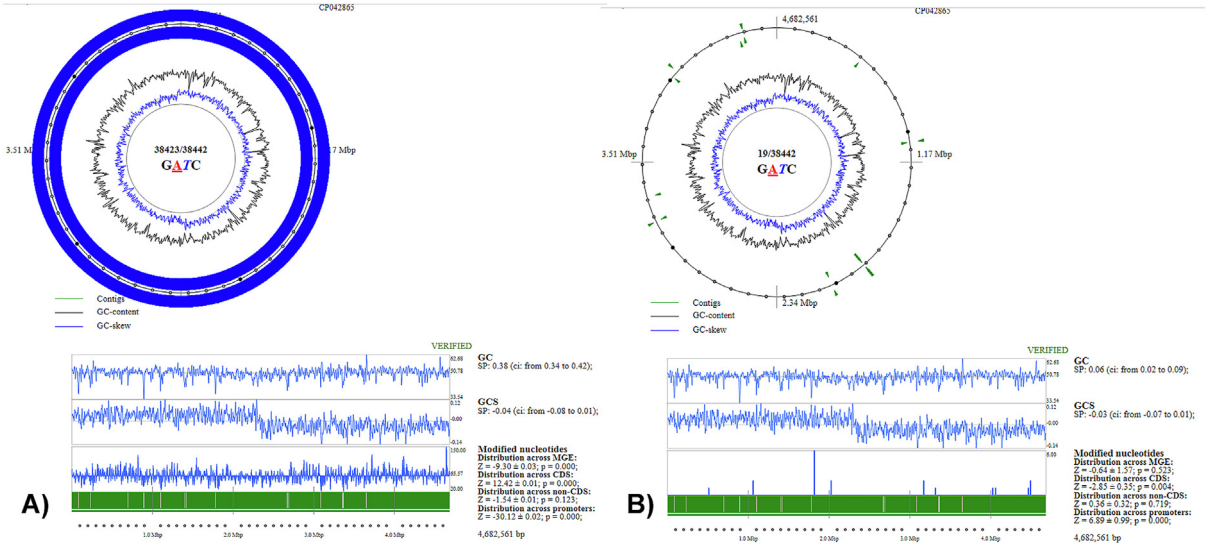

Figure 4. Circular maps showing the distribution of (A) methylated adenine residues associated with G*at*C canonical motifs and (B) adenine residues associated with GATC canonical motifs that remain unmethylated in the genome of *E. coli* ATCC BAA-39. Bacterial chromosomes are represented as circular lines, starting at the top of the circles, which correspond to chromosomal replication origins. Chromosomal coordinates are displayed alongside the circular graphs. The locations of modified bases are indicated by blue marks, while unmodified bases are represented by green marks, positioned outside or inside the circular graph to denote their presence on either the direct or reverse-complement strands of the chromosome. The statistical panels below the circular graphs present the distribution statistics for modified and unmodified bases, respectively.

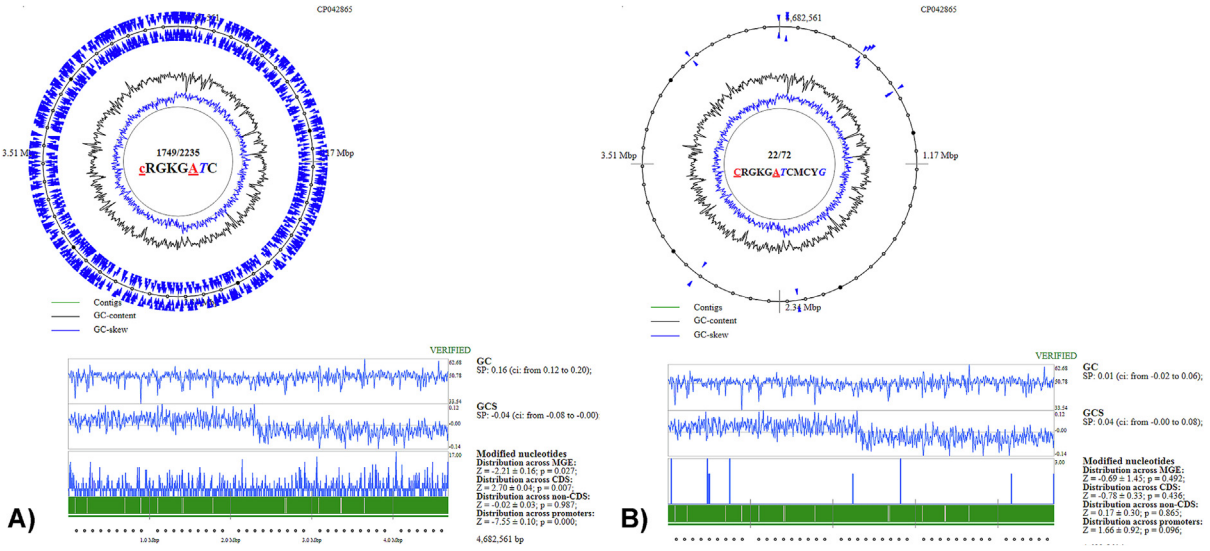

Figure 5. Circular maps showing the distribution of (A) methylated adenine and cytosine residues associated with *c*RGKG*at*C canonical motifs, and (B) *c*RGKG*at*CMCY*g* super-palindromes. The statistical panels below the circular graphs present the distribution statistics for modified bases associated with these two types of canonical motifs.

monality was the statistically significant abundance of methylated motifs in protein-coding genes.

Visualization of strand-specific orientation of canonical motifs in S. aureus genomes. The biological significance of epigenetic DNA modifications may vary depending on the location of modified bases on either direct or reverse-complement DNA strands. Specifically, modified bases can interfere with transcription only if they are located within promoter regions or gene bodies on the transcribed DNA strand. The program SWMM takes into account the strand location of modified bases when performing distribution statistics across coding, non-coding, and promoter regions. Furthermore, the program provides an option to display modified bases located on both strands or on one strand by setting the argument *--strand* to either *off* (default), *leading*, or *lagging*.

To exemplify these options, the distribution of methylated bases controlled by a type I RM system targeting GW*a*GNNNNNGA*t* motifs in *Staphylococcus aureus* 598 [CP082815] was analyzed using the command shown in Listing 3 (Suppl. Table S1).

The command shown in Listing 3 will display locations of modified adenine residues in the GW*a* part of the canonical motif across the chromosome, but only for direct inserts of the motif sequences. For the second run of the command, the argument *--cmap_motif* was reset to *GWaGNNNNNGAt,-1* to display the locations of methylated adenine residues at the reverse-complement GA*t* part of the canonical motif. The graphical outputs generated by these two commands are shown in Figure 6A and B, respectively. Changing the argument *--strand* to *lagging* will generate mirror reflections of the graphs shown in Figure 6.

The analysis demonstrated a highly biased distribution of methylated sites between the leading and lagging strands. On one half of the chromosome, clockwise from the replication origin to the replication terminus (the leading replichore), the GW*a* part of the canonical methylation motif is predominantly located on the direct strand, while the GA*t* part is located on the reverse-complement strand. This distribution is evident in the circular map and the statistical plot, and it is further supported by a strong positive correlation between GC skew and the frequency of GW*a* methylation, as well as a strong negative correlation with GA*t* methylation.

In contrast to Dam methylation in enterobacteria, which is abundant in protein-coding genes, an obvious avoidance of GW*a*GNNNNNGA*t* motifs was observed in coding and 120 bp TSC-upstream regions in the genome of *S. aureus* 598. Instead, these motifs were abundant in non-coding regions of the genome.

Visualization of distribution of methylated bases across different replicons. The program SWMM allows for statistical analysis to determine whether the frequency of DNA methylation is similar across different replicons, such as chromosomes and plasmids. For instance, the genome of *Haloferax volcanii* SVX82 consists of one chromosome and three plasmids. Nucleotide methylation in this genome occurs at adenine and cytosine residues associated with

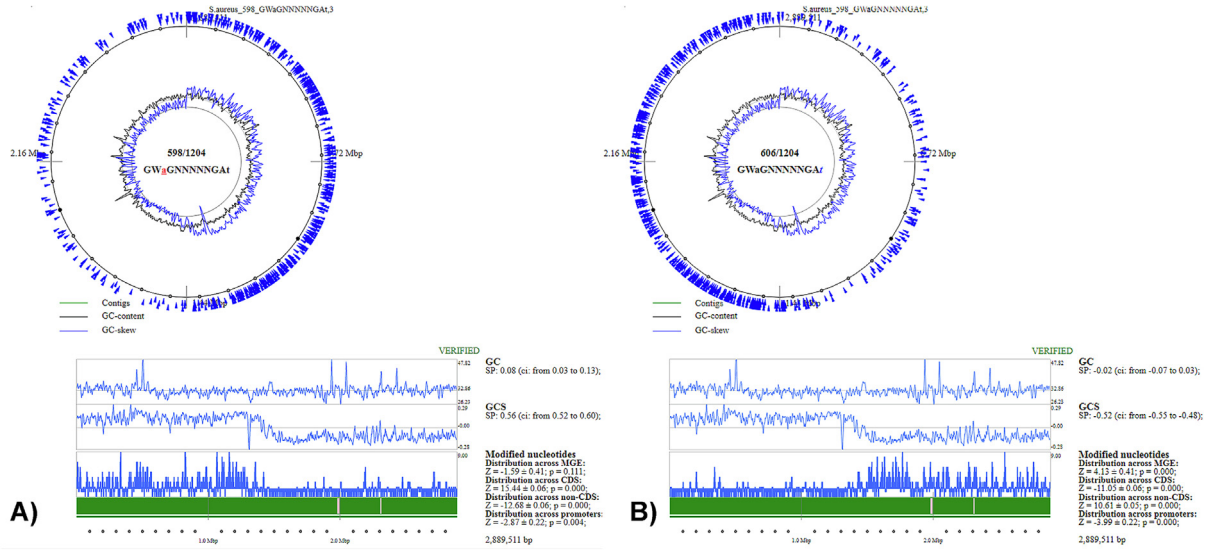

Figure 6. Circular maps showing the distribution of methylated adenine residues associated with (A) the GW*a* part of the canonical motif GW*a*GNNNNN, and (B) the GA*t* part of this motif in sequences located on the leading DNA strand of the of *S. aureus* 598 chromosome. The statistical panels below the circular graphs characterize the biased distribution of methylated adenine residues.

different canonical motifs [8]. The core structure of the cytosine-methylated motifs is the *c*TW*g* sequence, while the consensus motif for methylated adenines is *t*CG*a*. The replicons of the *H. volcanii* genome were combined into a single GBK file, and additional *contig* features were added to the *gbk.feature* list including the coordinates of the respective replicons.

The command in Listing 4 (Suppl. Table S1) was used to visualize the distribution of methylated cytosine residues across the chromosome and the plasmids of *H. volcanii* SVX82.

The graph resulting from the execution of Listing 4 is shown in Figure 7A. The distribution of methylated adenine residues was visualized using a similar command, where the argument *--cmap_motif* was set to *t*CG*a,1,-1.*

It was found that cytosine methylation at *c*TW*g* motifs was abundant in the first plasmid, but relatively rare in the chromosome and the other two plasmids (Figure 7A). Additionally, *c*TW*g* motifs were abundant in identified mobile genetic elements, non-coding regions, and the 120 bp TSC-upstream regions, but rare in protein-coding sequences. The latter observation may explain the negative correlation observed between the frequency of methylated cytosine residues and GC content.

Adenine methylation was less effective than cytosine methylation. Out of 103,990 adenine residues found within TCGA motifs, only 890 were methylated. In contrast to cytosine methylation, methylated adenine residues were more abundant in the chromosome and two plasmids, but rare in the second plasmid (Figure 7B). Methylated adenines were abundant in coding sequences, but rare in the 120 bp TSC-upstream regions.

Web-based access to SWMM

An online version of SWMM is available at https://begp.bi.up.ac.za. This web-based implementation of SWMM has certain restrictions on some parameter choices to conserve computational resources and ensure continued accessibility. For instance, the argument *--maximum_sites* is limited to a maximum of 10,000 modification sites per program run. Access to a standalone version is available at https://github.com/chrilef/BactEpiGenPro and allows for full parameter customization.

Discussion

The involvement of epigenetic modifications in bacterial genomes, such as adaptive evolution and genome maintenance [3–6,12], interactions within environmental microbial communities [8], and the development of antibiotic resistance and virulence traits [31], has become well recognized. However, the limited availability of computational tools for visualizing and statistically analyzing the distribution of epigenetically modified nucleotides in bacterial genomes has hindered further research in this field.

A significant step toward addressing this gap was made with the introduction of the BacMethy program [24]. In addition to visualizing genome methylation patterns, BacMethy enables the identification of methylated sites overlapping predicted

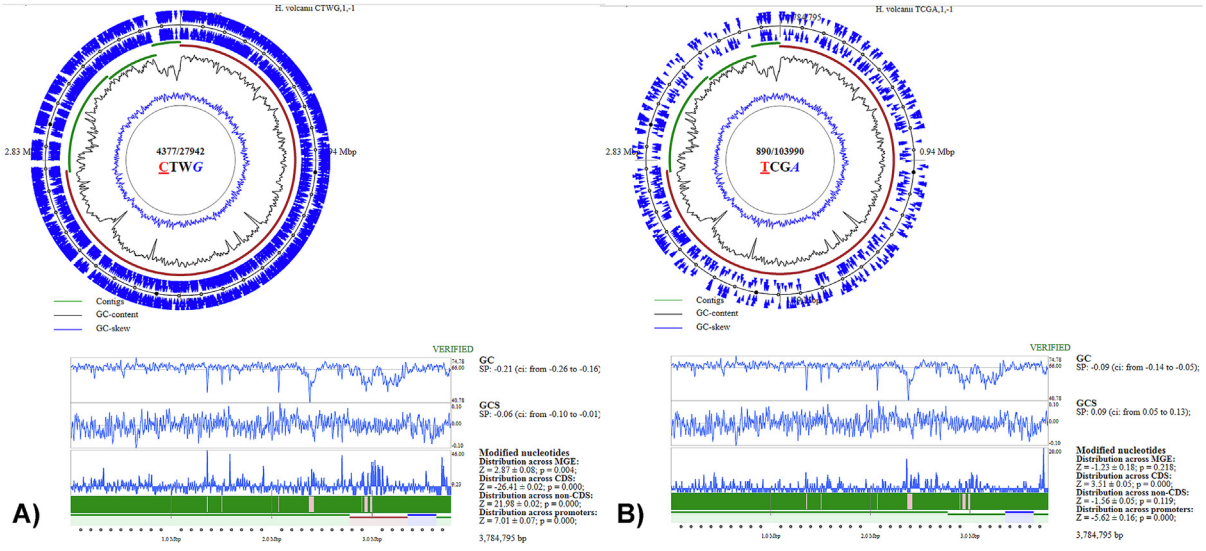

Figure 7. Circular maps of the distribution of (A) methylated adenine residues associated with *c*TW*g* canonical motifs, and (B) methylated cytosine residues associated with *t*CG*a* canonical motifs in the chromosome and plasmids of *H. volcanii* SVX82. The statistical panels below the circular graphs characterize the biased distribution of methylated bases across different replicons of this genome, represented by thin colored lines above the genomic coordinate scale. The color of a replicon line indicates the frequency of modified bases relative to expectations: green signifies that the frequency of modified bases falls within the expected range; red indicates a statistically significant abundance of modified bases in the replicon; blue denotes a rarity of modified bases in the respective replicon. A more detailed analysis of the distribution of modified bases across genomic replicons or contigs can be found in the program's text output.

transcription factor binding sites. This function allowed linking genome methylation with gene regulation.

With the SeqWord Motif Mapper (SWMM) program, we aimed to provide users with enhanced functionality for statistically evaluating the biased distribution of epigenetically modified bases across genomes, as well as greater flexibility in specifying DNA motifs to include or exclude based on their methylation status or strand location, whereas graphical output of the program are compatible with those produced by BacMethy.

SWMM allows users to search for methylated or unmethylated motifs of their design, as well as non-canonical epigenetically modified bases not associated with any motifs. Recent studies suggest that such non-canonical modifications may serve as markers for genes that are overexpressed or suppressed in different bacterial strains or under varying growth conditions [28,29].

By integrating the SeqWord Gene Island Sniffer program code [25], SWMM also enables the parallel identification of horizontally acquired genetic islands and their methylation levels. Additionally, our program verifies the locations of modified nucleotides predicted in GFF files using an internal BLASTN-based algorithm. This feature allows the generation of GFF files from multiple sequencing outputs by aligning them to the same reference sequence, facilitating comparative studies of epigenetic modification patterns across different isolates or conditions [8,29].

Results for user-defined motifs can be presented by the program in multiple formats, providing in-depth insights into methylated loci and sequence coverage. The program can be deployed on all popular operating systems, which have Python version 3.9 or higher installed.

To demonstrate the program's usage and practical utility, we conducted several studies on well-described bacterial strains, which yielded outcomes of practical interest. We demonstrated that Dam methylation, which is common in enterobacteria and many other microorganisms including phylogenetically distant Gram-positive bacteria, is more complex than previously known. In addition to adenine methylation at simple 4-base G*at*C palindromes, it also involves cytosine methylation at complex *c*RGKG*at*C motifs and even more complex *c*RGKG*at*CMCY*g* super-palindromes. Casadesús and Low (2006) reported that hemimethylation of GATC motifs in promoter regions is important in balancing gene expression and that blocking GATC methylation at certain locations is heritable and propagates phenotypic variations in bacterial populations [6]. This may explain the rarity of GATC motifs in the 120 bp TSC-upstream regions, as it helps avoid potential interference with gene regulation where it is not

needed. On the other hand, cytosine methylation in the cognate *c*RGKG*at*C motifs is more variable than adenine methylation and may potentially play a more significant role in gene regulation and the maintenance of transcriptional polymorphism in bacterial populations.

The abundance of methylated motifs in protein-coding genes and their potential to form hairpins in secondary mRNA structures suggest a possible role of methylation in these regions, not only in transcriptional regulation but also in post-transcriptional mRNA turnover – a phenomenon previously studied only in eukaryotes [32].

The strict orientation of canonical motifs along the replichores of the *S. aureus* genome was demonstrated. Semi-conserved repeated sequences aligned along the replichores of bacterial chromosomes are known as architecture-imparting sequences (AIMS), which are involved in the directional loading of FtsK translocases, proper chromosomal replication, and DNA repair [33]. Here, we report methylation of AIMS repeats in *S. aureus*, which may represent a previously unknown mechanism of AIMS functioning. Interestingly, the strain *S. aureus* 598 has another type I RM methylation at GG*a*NNNNNNN*t*CG [10], which does not exhibit such strand-biased distribution. The presence of two adenine methylation patterns, one associated with AIMS repeats and the other randomly distributed, has been reported in other genomes, such as MRSA *S. aureus* ATCC BAA-39 [CP033505] and *S. aureus* 150 [CP102945], whereas the genome of *S. aureus* 597/2 [CP082813] contains only AIMS-associated methylation [111]. The biased distribution of methylated bases across chromosomal replichores appears specific for *S. aureus* genomes, as the pattern was not observed in other tested microorganisms.

A biased distribution of methylated sites across the chromosome and plasmid was demonstrated in the archaeal genome of *H. volcanii* SVX82 [8]. Additionally, the program SWMM allows for the prediction of mobile genetic elements (MGEs) and enabled statistical analysis of the distribution of methylated sites across core parts of the chromosome and MGEs. It has often been observed that MGEs and plasmids have a smaller number of methylated sites. This avoidance strategy helps these MGEs evade host defenses, facilitating successful integration and propagation [34]. However, in the case of cytosine methylation in the genome of *H. volcanii* SVX82, a significantly higher frequency was observed in the large plasmid and MGE insertion regions in the chromosome. Another peculiarity of this methylation was that methylated cytosine residues were abundant in the 120 bp TSC-upstream regions. It can be hypothesized that this methylation plays a significant role in the plasmid lifecycle and its interaction with the host. Indeed, a transcriptionally active RM system was found on a plasmid [8]. The significant abundance of methylated cytosine residues within 120 bp TSC-upstream regions can be explained by the targeting of TAG stop codons by the *c*TW*g* canonical motifs in this archaeon. The majority of genes in the chromosome of *H. volcanii* SVX82 have TGA and TAA stop codons (1,725 and 637, respectively), whereas TAG stop codons are in the minority (511). This stop codon distribution may reflect a selective pressure to avoid methylation by the cytosine-specific MTase. The role of stop codon methylation in bacteria and archaea has not yet been studied.

The identification of nucleotide methylation patterns may further be useful in the reconstruction of genomes from metagenomics sequencing data as it would enable the grouping of assembled contigs originating from several closely related organisms.

Additionally, an example of the visualization and analysis of non-canonical nucleotide modifications in bacterial genomes was demonstrated. A recent study showed that non-canonical methylation is associated with differential gene expression in bacteria [1].

Conclusion

SeqWord Motif Mapper (SWMM) program was written in Python 3.9 and is freely available at https://github.com/chrilef/BactEpiGenPro. The program is also accessible as a Web-application at https://begp.bi.up.ac.za. SWMM provides an essential platform for analyzing bacterial methylation patterns, offering both visualization and statistical evaluation of modified nucleotide distributions. Through its implementation, we uncovered novel insights into bacterial epigenetics, including non-canonical methylation motifs, biased strand distributions, and epigenetic influences on genomic architecture. Notably, our findings highlight the complexity of Dam methylation, which extends beyond the well-characterized GATC palindromes to more intricate cytosine-methylated motifs, potentially influencing gene regulation and bacterial adaptability. Moreover, the strand-biased distribution of epigenetic marks in *Staphylococcus aureus* and the varied methylation patterns across plasmids and chromosomes in *Haloferax volcanii* suggest that epigenetic modifications contribute to horizontal gene transfer, genome stability, and adaptive evolution. The availability of SWMM as an open-source, platform-independent utility with both command-line and web-based interfaces ensures broad accessibility for researchers in microbiology, bioinformatics, and epigenetics. Future extensions of SWMM will focus on integrating additional sequencing technologies, estimation of statistical associations between alternative patterns of epigenetic modifications and differential gene expression, and expanding

the functional annotation of epigenetic modifications in epitranscriptomics.

CRediT authorship contribution statement

Christophe M.J. Lefebvre: Visualization, Software, Writing – original draft, Writing – review & editing, Validation, Investigation. **Rian E. Pierneef:** Writing – review & editing, Writing – original draft, Supervision, Software, Formal analysis, Data curation. **Oleg N. Reva:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Software, Writing – original draft, Writing – review & editing.

Funding

This project was funded by the National Research Foundation (NRF) of South Africa grant CPRR23030981722.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

ONR thanks the research team from the JSC Scientific Center for Anti-Infectious Drugs (SCAID), Almaty (Kazakhstan) led by Prof. Ilya S. Korotetskiy and the research team from the Institute of Polar Sciences, National Research Council (ISP – CNR), Messina (Italy) led by Prof. Michail M. Yakimov for previous collaboration resulted in publications cited in this paper and in publicly available SMRT PacBio reads used in this study to demonstrate functionality of the SeqWord Motif Mapper (SWMM) program.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.jmb.2025.169307.

Received 26 April 2025;
Accepted 22 June 2025;
Available online 25 June 2025

Keywords:
software;
Python;
biostatistics;
epigenetics;
methylomics

References

[1]. Beaulaurier, J., Schadt, E.E., Fang, G., (2019). Deciphering bacterial epigenomes using modern sequencing technologies. *Nature Rev. Genet.* **20** (3), 157–172. https://doi.org/10.1038/s41576-018-0081-3.

[2]. Kino, K., Hirao-Suzuki, M., Morikawa, M., Sakaga, A., Miyazawa, H., (2017). Generation, repair and replication of guanine oxidation products. *Genes Environ.* **39**, 21. https://doi.org/10.1186/s41021-017-0081-0.

[3]. Reisenauer, A., Kahng, L.S., McCollum, S., Shapiro, L., (1999). Bacterial DNA methylation: a cell cycle regulator?. *J. Bacteriol.* **181** (17), 5135–5139. https://doi.org/10.1128/JB.181.17.5135-5139.1999.

[4]. Riva, A., Delorme, M.O., Chevalier, T., Guilhot, N., Hénaut, C., Hénaut, A., (2004). Characterization of the GATC regulatory network in *E. coli. BMC Genomics* **5** (1), 48. https://doi.org/10.1186/1471-2164-5-48.

[5]. Sánchez-Romero, M.A., Casadesús, J., (2020). The bacterial epigenome. *Nature Rev. Microbiol.* **18** (1), 7–20. https://doi.org/10.1038/s41579-019-0286-2.

[6]. Casadesús, J., Low, D., (2006). Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.* **70** (3), 830–856. https://doi.org/10.1128/MMBR.00016-06.

[7]. Bheemanaik, S., Reddy, Y.V., Rao, D.N., (2006). Structure, function and mechanism of exocyclic DNA methyltransferases. *Biochem. J.* **399** (2), 177–190. https://doi.org/10.1042/BJ20060854.

[8]. Reva, O.N., La Cono, V., Crisafi, F., Smedile, F., Mudaliyar, M., Ghosal, D., Giuliano, L., Krupovic, M., Yakimov, M.M., (2024). Interplay of intracellular and trans-cellular DNA methylation in natural archaeal consortia. *Environ. Microbiol. Rep.* **16**, (2)e13258 https://doi.org/10.1111/1758-2229.13258.

[9]. Roberts, R.J., Belfort, M., Bestor, T., Bhagwat, A.S., Bickle, T.A., Bitinaite, J., Blumenthal, R.M., Degtyarev, S. K., Dryden, D.T., Dybvig, K., Firman, K., Gromova, E.S., Gumport, R.I., Halford, S.E., Hattman, S., Heitman, J., Hornby, D.P., Janulaitis, A., Jeltsch, A., Josephsen, J., Kiss, A., Klaenhammer, T.R., Kobayashi, I., Kong, H., Krüger, D.H., Lacks, S., Marinus, M.G., Miyahara, M., Morgan, R.D., Murray, N.E., Nagaraja, V., Piekarowicz, A., Pingoud, A., Raleigh, E., Rao, D.N., Reich, N., Repin, V.E., Selker, E.U., Shaw, P.C., Stein, D.C., Stoddard, B. L., Szybalski, W., Trautner, T.A., Van Etten, J.L., Vitor, J. M., Wilson, G.G., Xu, S.Y., (2003). A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* **31** (7), 1805–1812. https://doi.org/10.1093/nar/gkg274.

[10]. Korotetskiy, I.S., Shilov, S.V., Kuznetsova, T., Kerimzhanova, B., Korotetskaya, N., Ivanova, L., Zubenko, N., Parenova, R., Reva, O.N., (2023). Analysis of whole-genome sequences of pathogenic Gram-positive and Gram-negative isolates from the same hospital environment to investigate common evolutionary trends associated with horizontal gene exchange, mutations and DNA methylation patterning. *Microorganisms* **11** (2), 323. https://doi.org/10.3390/microorganisms11020323.

[11]. Passeri, I., Cangioli, L., Fondi, M., Mengoni, A., Fagorzi, C., (2025). The complex epigenetic panorama in the multipartite genome of the nitrogen-fixing bacterium *Sinorhizobium meliloti*. *Genome Biol. Evol.* **17**, (1) evae245 https://doi.org/10.1093/gbe/evae245.

[12]. Peterson, S.N., Reich, N.O., (2008). Competitive Lrp and Dam assembly at the *pap* regulatory region: implications for mechanisms of epigenetic regulation. *J. Mol. Biol.* **383** (1), 92–105. https://doi.org/10.1016/j.jmb.2008.07.086.

[13]. Reva, O.N., Swanevelder, D.Z.H., Mwita, L.A., Mwakilili, A.D., Muzondiwa, D., Joubert, M., Chan, W.Y., Lutz, S., Ahrens, C.H., Avdeeva, L.V., Kharkhota, M.A., Tibuhwa, D., Lyantagaye, S., Vater, J., Borriss, R., Meijer, J., (2019). Genetic, epigenetic and phenotypic diversity of four *Bacillus velezensis* strains used for plant protection or as probiotics. *Front. Microbiol.* **10**, 2610. https://doi.org/10.3389/fmicb.2019.02610.

[14]. Vasilchenko, N.G., Prazdnova, E.V., Lewitin, E., (2022). Epigenetic mechanisms of gene expression regulation in bacteria of the genus *Bacillus*. *Russ. J. Genet.* **58** (1), 1–9.

[15]. Dalia, A.B., Lazinski, D.W., Camilli, A., (2013). Characterization of undermethylated sites in *Vibrio cholerae*. *J. Bacteriol.* **195** (10), 2389–2399. https://doi.org/10.1128/JB.02112-12.

[16]. Kahramanoglou, C., Prieto, A.I., Khedkar, S., Haase, B., Gupta, A., Benes, V., Fraser, G.M., Luscombe, N.M., Seshasayee, A.S., (2012). Genomics of DNA cytosine methylation in *Escherichia coli* reveals its role in stationary phase transcription. *Nature Commun.* **3**, 886. https://doi.org/10.1038/ncomms1878.

[17]. Bart, A., van Passel, M.W., van Amsterdam, K., van der Ende, A., (2005). Direct detection of methylation in genomic DNA. *Nucleic Acids Res.* **33** (14), e124.

[18]. Fang, G., Munera, D., Friedman, D.I., Mandlik, A., Chao, M.C., Banerjee, O., Feng, Z., Losic, B., Mahajan, M.C., Jabado, O.J., Deikus, G., Clark, T.A., Luong, K., Murray, I.A., Davis, B.M., Keren-Paz, A., Chess, A., Roberts, R.J., Korlach, J., Turner, S.W., Kumar, V., Waldor, M.K., Schadt, E.E., (2012). Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nature Biotechnol.* **30** (12), 1232–1239. https://doi.org/10.1038/nbt.2432. Epub 2012 Nov 8. Erratum in: Nature Biotechnol. 2013;31(6):566..

[19]. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J., Turner, S.W., (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods* **7** (6), 461–465. https://doi.org/10.1038/nmeth.1459.

[20]. Couturier, M., Lindås, A.C., (2018). The DNA methylome of the hyperthermoacidophilic crenarchaeon *Sulfolobus acidocaldarius*. *Front. Microbiol.* **9**, 137. https://doi.org/10.3389/fmicb.2018.00137.

[21]. Tourancheau, A., Mead, E.A., Zhang, X.S., Fang, G., (2021). Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nature Methods* **18** (5), 491–498. https://doi.org/10.1038/s41592-021-01109-3.

[22]. Kulkarni, O., Mathew, R.J., Zaveri, L., Jana, R., Singh, N. K., Ara, S., Tallapaka, K.B., Sowpati, D.T., (2024). Comprehensive benchmarking of tools for nanopore-based detection of DNA methylation. *bioRxiv*. 2024-11.

[23]. Roberts, R.J., Vincze, T., Posfai, J., Macelis, D., (2023). REBASE: a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **51** (D1), D629–D630. https://doi.org/10.1093/nar/gkac975.

[24]. Liu, J.H., Zhang, Y., Zhou, N., He, J., Xu, J., Cai, Z., Yang, L., Liu, Y., (2024). Bacmethy: a novel and convenient tool for investigating bacterial DNA methylation pattern and their transcriptional regulation effects. *iMeta* **3** (3), e186.

[25]. Bezuidt, O., Lima Mendez, G., Reva, O.N., (2009). SEQWord gene island sniffer: a program to study the lateral genetic exchange among bacteria. *World Acad. Sci. Eng. Technol.* **58**, 1169–1174.

[26]. Reva, O.N., Larisa, S.A., Mwakilili, A.D., Tibuhwa, D., Lyantagaye, S., Chan, W.Y., Lutz, S., Ahrens, C.H., Vater, J., Borriss, R., (2020). Complete genome sequence and epigenetic profile of *Bacillus velezensis* UCMB5140 used for plant and crop protection in comparison with other plant-associated *Bacillus* strains. *Appl. Microbiol. Biotechnol.* **104** (17), 7643–7656. https://doi.org/10.1007/s00253-020-10767-w.

[27]. Korotetskiy, I.S., Jumagaziyeva, A.B., Shilov, S.V., Kuznetsova, T.V., Myrzabayeva, A.N., Iskakbayeva, Z. A., Ilin, A.I., Joubert, M., Taukobong, S., Reva, O.N., (2021). Transcriptomics and methylomics study on the effect of iodine-containing drug FS-1 on *Escherichia coli* ATCC BAA-196. *Future Microbiol.* **16**, 1063–1085. https://doi.org/10.2217/fmb-2020-0184.

[28]. Korotetskiy, I., Shilov, S., Kuznetsova, T., Zubenko, N., Ivanova, L., Reva, O.N., (2025). Epigenetic background of lineage-specific genome expression landscapes of four *Staphylococcus aureus* hospital isolates. *PLoS One* **20**, (4) e0322006 https://doi.org/10.1371/journal.pone.0322006.

[29]. Smedile, F., Denaro, R., Crisafi, F., Giosa, D., D'Auria, G., Ferrer, M., Rosselli, R., Staege, M.S., Yakimov, M.M., Giuliano, L., Reva, O.N., (2025). Integrated analysis of gene expression, protein synthesis, and epigenetic modifications in *Alcanivorax borkumensis* SK2 under iron limitation. *Environ. Microbiol. Rep.* **17**, (3)e70106 https://doi.org/10.1111/1758-2229.70106.

[30]. Li, H., (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34** (18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

[31]. Shell, S.S., Prestwich, E.G., Baek, S.H., Shah, R.R., Sassetti, C.M., Dedon, P.C., Fortune, S.M., (2013). DNA methylation impacts gene expression and ensures hypoxic survival of *Mycobacterium tuberculosis*. *PLoS Pathog.* **9**, (7)e1003419 https://doi.org/10.1371/journal.ppat.1003419.

[32]. Peer, E., Rechavi, G., Dominissini, D., (2017). Epitranscriptomics: regulation of mRNA metabolism through modifications. *Curr. Opin. Chem. Biol.* **41**, 93–98. https://doi.org/10.1016/j.cbpa.2017.10.008.

[33]. Hendrickson, H.L., Barbeau, D., Ceschin, R., Lawrence, J.G., (2018). Chromosome architecture constrains horizontal gene transfer in bacteria. *PLoS Genet.* **14**, e1007421. https://doi.org/10.1371/journal.pgen.1007421.

[34]. Shaw, L.P., Rocha, E.P.C., MacLean, R.C., (2023). Restriction-modification systems have shaped the evolution and distribution of plasmids across bacteria. *Nucleic Acids Res.* **51** (13), 6806–6818. https://doi.org/10.1093/nar/gkad452.