Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 22 August2026
Revised 24 September 2026
Accepted 05 October 2026

Natural Resources for Human Health 2026, Vol. 6 (16s): 1028–1040

KEYWORDS:
Bacteriophage, multidrug-resistant Escherichia coli, whole-genome sequencing, MASH, Pharokka, terL, comparative genomics, phage genomics

Whole-Genome Characterization and Comparative Genomic Analysis of Three Bacteriophages Active against Multidrug-Resistant Escherichia coli

Mukesh Kumar Singh[1*], Ashutosh Rawat[2], Asim Kumar Singh[3]

[1*]PhD Scholar, Department of Microbiology, Santosh Medical College and Hospital, Santosh Deemed to be University, Ghaziabad, Uttar Pradesh, India
Orcid ID: https://orcid.org/0009-0003-1680-7378 , Email: mukudeos@gmail.com
[2]Professor and Head, Department of Microbiology, Santosh Medical College and Hospital, Santosh Deemed to be University, Ghaziabad, Uttar Pradesh, India
[3]Professor and Head, Department of Microbiology, Heritage Institute of Medical Sciences, Varanasi, Uttar Pradesh, India

ABSTRACT: **Background**: Multidrug-resistant (MDR) Escherichia coli is an important clinical challenge, and bacteriophages are being investigated as targeted antibacterial agents. Whole-genome sequencing is essential for defining phage genome organization, taxonomic relationships and genomic features relevant to phage biology and safety.

Methods: Three purified bacteriophages active against MDR E. coli, isolated from environmental water samples and designated Phage-1 (P1; sequencing sample 1A), Phage-2 (P2; sample 2A) and Phage-3 (P3; sample 3A), were subjected to Illumina paired-end whole-genome sequencing. Raw-read quality was assessed using FastQC and MultiQC, and reads were trimmed using Trim Galore. Genomes were assembled using Unicycler and annotated using Pharokka. MASH/INPHARED was used for comparative genomic identification. Predicted lysis-associated genes and CARD/VFDB hits were assessed, and terminase large-subunit (terL)-based phylogenetic analysis was performed.

Results: Single-contig assemblies of 50,343 bp, 31,564 bp and 39,699 bp were obtained for P1, P2 and P3, respectively, with GC contents of 44.395%, 51.384% and 49.777%. MASH identified Escherichia phage vB_Eco_mar001J1, Escherichia phage 11W and Escherichia phage K1F as the closest database references for P1, P2 and P3, respectively, each with a MASH distance of 0.0 and 1000/1000 matching hashes. In the terL-based phylogeny, P1 clustered with vB_Eco_mar001J1, P2 with phage 11W and other P2-like (Peduovirus) phages, and P3 with Escherichia phage MirsaeidiM_Esher2. Lysis-associated functions were detected in all three genomes. No CARD antimicrobial-resistance-gene hits or VFDB virulence-factor hits were detected. An integrase annotation was identified in P2.

Conclusion: The three bacteriophages showed distinct genome sizes and functional organizations and were closely related to previously described E. coli phages. The absence of detected CARD/VFDB hits and integration-associated genes supports further evaluation of P1 and P3, whereas the integrase annotation and P2-like affiliation of P2 suggest a potentially temperate lifestyle that limits its therapeutic suitability. The genomic data provide a foundation for further biological and taxonomic characterization.

Natural Resources for Human Health
eISSN: 2583-1194

1. INTRODUCTION

The global increase in antimicrobial resistance has reduced the effectiveness of commonly used antibacterial agents and has increased interest in alternative approaches for controlling resistant bacterial infections. [1] Escherichia coli is a frequent cause of community- and healthcare-associated infections and can acquire resistance to multiple antimicrobial classes. [1,2] Bacteriophages, viruses that infect bacteria, have therefore received renewed attention because of their host specificity and capacity to lyse susceptible bacterial cells. [3]

Phenotypic phage characterization can establish host range and lytic activity, but genomic characterization is required to define the genetic basis of phage replication, structural assembly and host-cell lysis and to identify genomic features that may influence suitability for further development. Modern sequencing workflows allow phage genomes to be assembled and annotated using reproducible bioinformatic pipelines. Guidance for phage genome annotation emphasizes careful assembly, annotation and interpretation before genome submission. [4]

Comparative genomic approaches can additionally establish relationships between newly characterized phages and previously deposited genomes. Alignment-free methods such as MASH permit rapid comparison of genome sketches and identification of closely related reference genomes. [5] Genome annotation tools developed specifically for bacteriophages, such as Pharokka, facilitate standardized prediction and functional classification of phage coding sequences. [6]

In the present study, three bacteriophages active against MDR E. coli were subjected to whole-genome sequencing and comparative genomic analysis. The objective was to characterize their genome size and composition, assembly quality, functional annotation, lysis-associated genes, antimicrobial-resistance and virulence-associated database hits, comparative genomic relationships and terL-based phylogenetic placement.

2. MATERIALS AND METHODS

2.1 Bacteriophage isolates and genomic DNA

Three bacteriophages active against MDR E. coli, previously isolated from environmental water samples (pond water, sewage and Ganga river water) using clinical MDR E. coli isolates as propagation hosts, were selected for whole-genome sequencing and designated Phage-1 (P1; sample 1A), Phage-2 (P2; sample 2A) and Phage-3 (P3; sample 3A). High-titre phage lysates were submitted to Biokart Genomics Lab (Biokart India Pvt. Ltd., Bengaluru, India) for DNA extraction, sequencing and bioinformatic analysis under project ID BKWGS2596. [7-9]

2.2 Phage DNA extraction and quality control

Phage genomic DNA was extracted from the lysates using the Xploregen DNA Extraction Kit according to the manufacturer's instructions. DNA concentration was measured on a Qubit 3 Fluorometer (Thermo Fisher Scientific, USA) using the dsDNA High Sensitivity assay, purity was checked by NanoDrop spectrophotometry (an A260/A280 ratio of approximately 1.8–2.0 was considered acceptable), and integrity was assessed by agarose gel electrophoresis. [10]

2.3 Library preparation and sequencing

Sequencing libraries were prepared from 100 ng of genomic DNA per sample. DNA was enzymatically fragmented to a target size of 200–300 bp, followed by end repair and A-tailing. Loop adapters were ligated and cleaved with Uracil-Specific Excision Reagent (USER) enzyme, and the ligated products were purified with AMPure XP magnetic beads (Beckman Coulter, USA). Libraries were enriched by six PCR cycles using NEBNext Ultra II Q5 Master Mix (New England Biolabs, USA) with the Illumina universal primer and sample-specific index primers, purified with AMPure beads and eluted in 15 μL of 0.1× TE buffer. Libraries were quantified with the Qubit dsDNA HS assay, and fragment-size distribution was assessed on an Agilent 2100 Bioanalyzer using a DNA 7500 chip (Agilent Technologies, USA) and by agarose gel electrophoresis against a 100-bp DNA ladder. Libraries were pooled in equimolar amounts and sequenced on an Illumina platform with 2 × 150-bp paired-end chemistry, and reads were demultiplexed using sample-specific index sequences.

2.4 Read quality control and trimming

Raw reads were assessed with FastQC v0.12.1, and per-sample reports were combined with MultiQC v1.32. [11,12] Quality assessment included per-base sequence quality, per-sequence quality, per-tile sequence quality, base composition, GC-content

distribution, per-base N content, read-length distribution, sequence duplication, overrepresented sequences and adapter content. Adapter sequences and low-quality bases (Phred score <20) were removed with Trim Galore v0.6.10, and reads shorter than 20 bp after trimming were discarded. [24]

2.5 Genome assembly

Trimmed reads were assembled de novo with Unicycler v0.5.1, which, for Illumina-only data, optimizes assemblies generated by SPAdes v4.2.0. [13,25] Assembly statistics including total genome length, number of contigs, N50, L50, ambiguous bases and gaps were recorded.

2.6 Genome annotation

Genome annotation was performed using Pharokka v1.7.4 (database v1.7.4). Predicted coding sequences were categorized into functional groups including DNA/RNA/nucleotide metabolism, head and packaging, lysis, tail, transcriptional regulation, integration/excision, connector proteins, auxiliary/host-takeover functions and unknown function. Pharokka is designed specifically for standardized bacteriophage genome annotation. [6] Within the Pharokka workflow, coding sequences are predicted with PHANOTATE, functional annotation is assigned against the PHROG database using MMseqs2, tRNAs and tmRNAs are identified with tRNAscan-SE and ARAGORN, CRISPR arrays are detected with a CRT-based tool, and annotation outputs are screened against CARD and VFDB. [14-21]

2.7 Comparative genomic analysis

MASH v2.3 was used to compare the assembled phage genomes against the INPHARED phage database. The closest reference genome, reported MASH distance and matching-hash information were recorded. MASH is an alignment-free genome comparison method based on MinHash sketches. [5,22]

2.8 Phylogenetic analysis

The terminase large-subunit (terL) gene (1A_CDS_0018, 2A_CDS_0002 and 3A_CDS_0049) was used as the phylogenetic marker. Each terL nucleotide sequence was queried against the NCBI nucleotide (nr/nt) database using BLASTN (accessed 2 September 2026), and neighbour-joining trees of the closest hits were constructed from the BLAST pairwise alignments with a maximum sequence difference of 0.75. [23]

2.9 Nucleotide composition and genome visualization

Genome length and nucleotide composition were determined from the assembled genome sequences. Circular genome maps were generated to visualize genome organization; these maps were used for visualization only and were not taken as evidence of physical genome circularity.

3. Results

3.1 DNA quality and sequencing overview

Whole-genome sequencing was performed for P1, P2 and P3. DNA concentrations were 148 ng/μL, 130 ng/μL and 163 ng/μL, respectively; all three samples showed a single high-molecular-weight band without visible degradation and passed the initial quality-control assessment (Table 1; Figure 1). All three libraries showed a single discrete band of the expected size.

Table 1. DNA concentration and initial quality-control status of phage genomic DNA.

Sample	Phage	DNA concentration (ng/μL)	Initial QC
1A	P1	148	Pass
2A	P2	130	Pass
3A	P3	163	Pass

Natural Resources for Human Health
eISSN: 2583-1194

Genomic DNA gel image as visualised in gel electrophoresis



DNA Library QC gel image as visualised in gel electrophoresis



Figure 1. Agarose gel electrophoresis of phage genomic DNA (upper panel) and prepared sequencing libraries (lower panel) for P1 (lane 1), P2 (lane 2) and P3 (lane 3).

Sequencing generated 8,908,630, 11,192,740 and 13,746,512 raw reads for P1, P2 and P3, respectively, of which 8,908,092, 11,192,112 and 13,745,764 (>99.99%) were retained after trimming. The MultiQC/FastQC report contained six read files (R1 and R2 for each phage; Table 2). All files had an average/median read length of 151 bp. Per-base sequence quality and per-sequence quality modules passed for all files. Per-tile sequence quality was flagged for the P1 R1 file and passed for the other five files. Per-sequence GC-content, sequence duplication and adapter-content modules were flagged across the datasets. Such flags are common in phage sequencing data, where very high read depth over a small genome inflates duplication levels, and did not indicate a failed sequencing run.

Table 2. Raw-read statistics from FastQC/MultiQC.

Read file	Reads (M)	GC (%)	Avg/median length (bp)	Duplication (%)
P1 R1	4.454	61	151	88.56
P1 R2	4.454	61	151	88.49
P2 R1	5.596	59	151	56.85
P2 R2	5.596	59	151	56.76
P3 R1	6.873	62	151	56.55
P3 R2	6.873	63	151	56.12



Figure 2. Per-base sequence quality of raw sequencing reads for Phage-1 (1A), Phage-2 (2A) and Phage-3 (3A) (MultiQC/FastQC).

3.2 Genome assembly characteristics

All three phage genomes were assembled into single contigs. P1 had the largest assembled genome (50,343 bp), followed by P3 (39,699 bp) and P2 (31,564 bp). Because each assembly consisted of a single contig, the N50 equalled the complete assembled length and L50 was 1 for all three phages. P1 and P3 contained no ambiguous bases or reported gaps, whereas P2 contained 269 ambiguous N bases and 85 reported gaps.

Table 3. Assembly statistics of P1, P2 and P3.

Parameter	**P1**	**P2**	**P3**
Genome length (bp)	50,343	31,564	39,699
Contigs	1	1	1
N50 (bp)	50,343	31,564	39,699
L50	1	1	1
N bases	0	269	0
Reported gaps	0	85	0

Natural Resources for Human Health
eISSN: 2583-1194

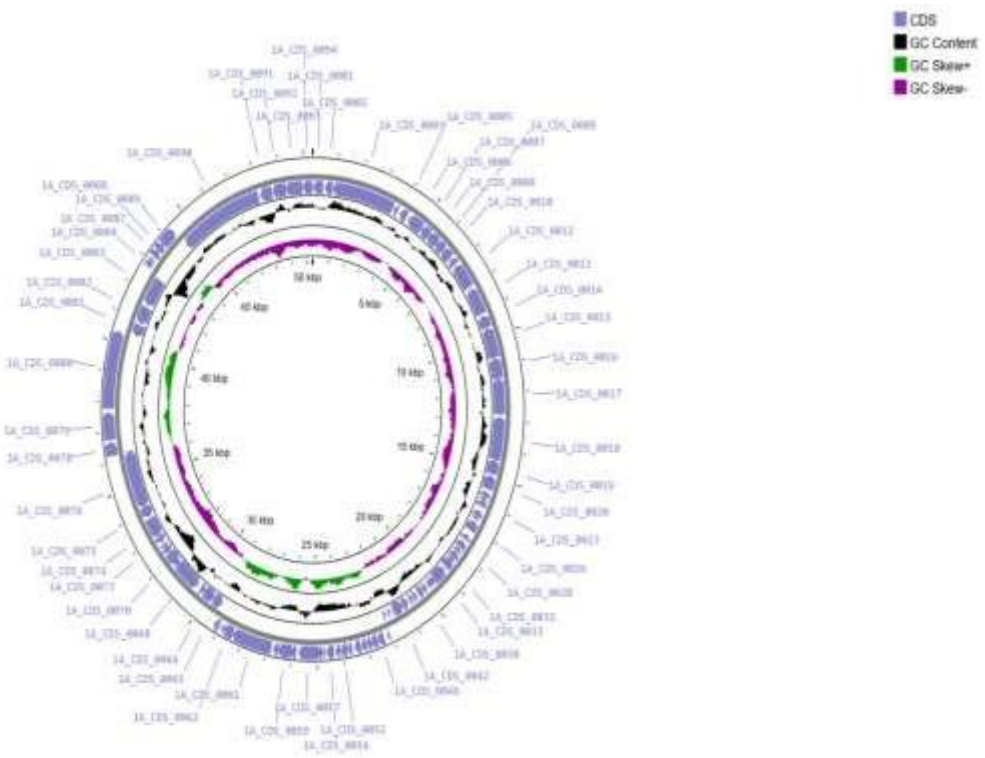


Figure 3. Circular genome map of Phage-1 (sample 1A), showing the organization of predicted genomic features across the 50,343-bp assembled genome

3.3 Genome composition and coding density

The assembled genomes showed GC contents of 44.395% for P1, 51.384% for P2 and 49.777% for P3. Coding density was 93.90%, 95.76% and 92.32%, respectively. P2 therefore had the highest coding density among the three assemblies.

Table 4. Genome composition and coding density.

Parameter	P1	P2	P3
Genome size (bp)	50,343	31,564	39,699
GC content (%)	44.395	51.384	49.777
Coding density (%)	93.90	95.76	92.32
Translation table	11	11	11

3.4 Comparative genomic analysis by MASH/INPHARED

MASH-based comparison against the INPHARED database identified Escherichia phage vB_Eco_mar001J1, Escherichia phage 11W and Escherichia phage K1F as the closest references for P1, P2 and P3, respectively (Table 5). The MASH distance was 0.0 with 1000/1000 matching hashes for all three phages, indicating that each study phage is very closely related at the whole-genome level to a previously described phage. These values do not by themselves establish that the study isolates are identical to the reference genomes.

Table 5. Closest reference phages identified by MASH/INPHARED.

Study phage	Closest reference phage	MASH distance	Matching hashes	INPHARED classification
P1	Escherichia phage vB_Eco_mar001J1	0.0	1000/1000	Warwickvirus; Drexlerviridae; Tempevirinae
P2	Escherichia phage 11W	0.0	1000/1000	Peduovirus; Peduoviridae
P3	Escherichia phage K1F	0.0	1000/1000	Kayfunavirus; Autographiviridae

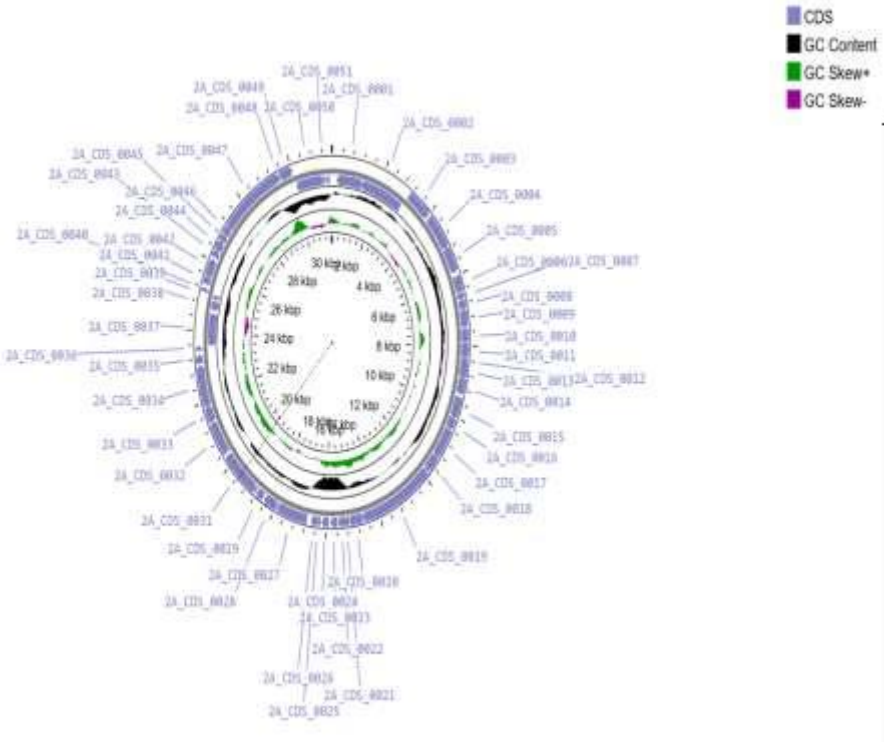


Figure 4. Circular genome map of Phage-2 (sample 2A), showing the organization of predicted genomic features across the 31,564-bp assembled genome.

3.5 Genome annotation and functional categories

Pharokka identified 94, 53 and 54 predicted coding sequences in P1, P2 and P3, respectively. A substantial fraction of predicted proteins remained in the unknown-function category, with 52, 18 and 24 annotations, respectively.

Table 6. Functional categories of predicted genes (Pharokka).

Functional category	P1	P2	P3
CDS	94	53	54
Connector	3	4	1
DNA/RNA/nucleotide metabolism	11	3	10
Head and packaging	7	5	8
Integration/excision	0	1	0
Lysis	3	4	3

Moron/auxiliary/host takeover	2	0	3
Other	4	1	2
Tail	11	15	3
Transcription regulation	1	2	0
Unknown function	52	18	24
tRNAs	0	0	0
CRISPRs	0	0	0
tmRNAs	0	0	0
CARD AMR hits	0	0	0
VFDB virulence hits	0	0	0

3.6 Lysis-associated genes

Lysis-associated annotations were identified in all three phage genomes. P1 contained an Rz-like spanin, endolysin and holin. P2 contained a holin, endolysin and two Rz-like spanin annotations. P3 contained an amidase, holin and Rz-like spanin.

Table 7. Lysis-associated genes identified in the three phage genomes.

Phage	Gene/feature	Predicted function
P1	1A_CDS_0065	Rz-like spanin
P1	1A_CDS_0066	Endolysin
P1	1A_CDS_0067	Holin
P2	2A_CDS_0008	Holin
P2	2A_CDS_0010	Endolysin
P2	2A_CDS_0011	Rz-like spanin
P2	2A_CDS_0012	Rz-like spanin
P3	3A_CDS_0022	Amidase
P3	3A_CDS_0046	Holin
P3	3A_CDS_0048	Rz-like spanin

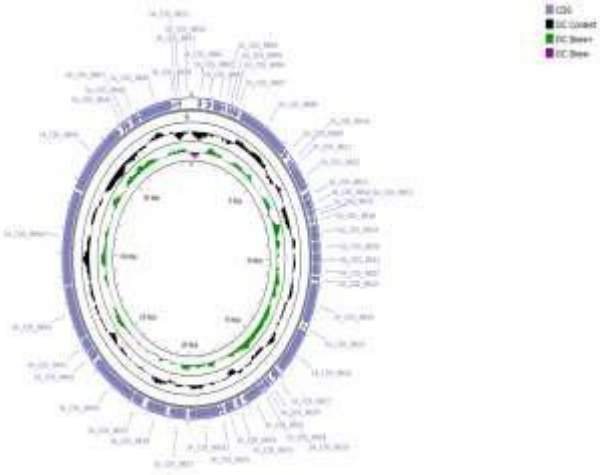


Figure 5. Circular genome map of Phage-3 (sample 3A), showing the organization of predicted genomic features across the 39,699-bp assembled genome.

3.7 Integration/excision annotation and predicted lifestyle

No integration/excision-associated annotation was identified in P1 or P3. In contrast, P2 contained one annotated integrase (2A_CDS_0037), classified under integration and excision. Accordingly, P2 was not described as unequivocally strictly lytic on the basis of genomic annotation alone. The integrase finding should be interpreted as a genomic feature requiring consideration in downstream phage-development decisions.

Table 8. Integration/excision annotations and lifestyle interpretation.

Phage	Integration/excision annotation	Interpretation
P1	0	No integration/excision-associated annotation detected
P2	1 integrase (2A_CDS_0037)	Lifestyle not classified unequivocally as strictly lytic from WGS alone
P3	0	No integration/excision-associated annotation detected

3.8 Antimicrobial-resistance and virulence screening

No antimicrobial-resistance-gene hits were identified against CARD and no virulence-factor hits were identified against VFDB in any of the three annotated phage genomes. No tRNA genes or CRISPR arrays were identified in the annotation output. These findings provide supportive genomic information within the scope of the databases and annotation pipeline used; absence of database hits does not prove the absolute absence of all biologically relevant genes.

3.9 terL-based phylogenetic analysis

In the terL-based phylogenetic analysis, P1 clustered with Escherichia virus vB_Eco_mar001J1 sequences (Figure 6), and P2 clustered with Escherichia phage 11W and other P2-like (Peduovirus) phages (Figure 7); both placements were consistent with the MASH results. The terL sequence of P3 clustered most closely with Escherichia phage MirsaeidiM_Esher2 (Figure 8); phage K1F was not among the BLASTN hits included in this tree. Branch-support (bootstrap) values were not calculated for these distance trees.

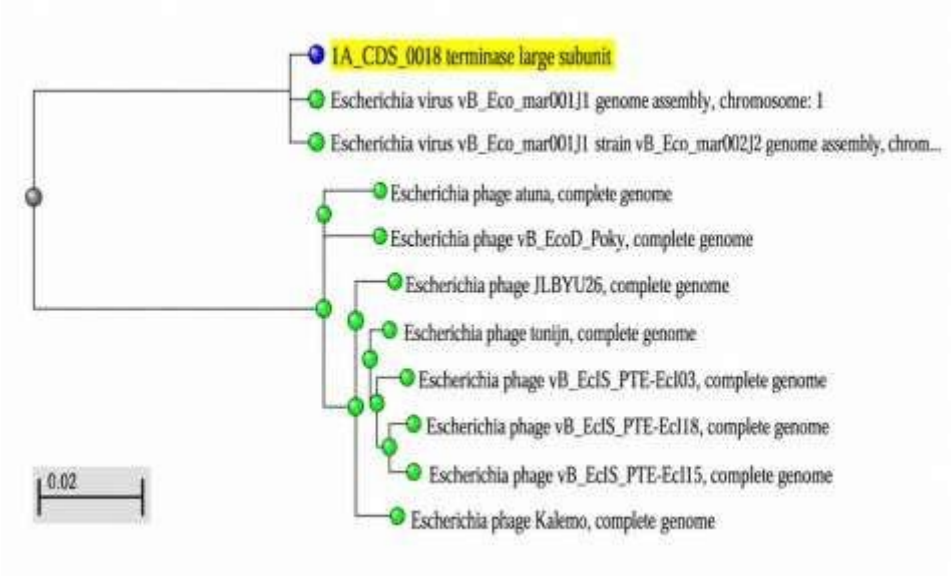


Figure 6. Terminase large-subunit (terL)-based phylogenetic tree of Phage-1 (1A) and related phage sequences.

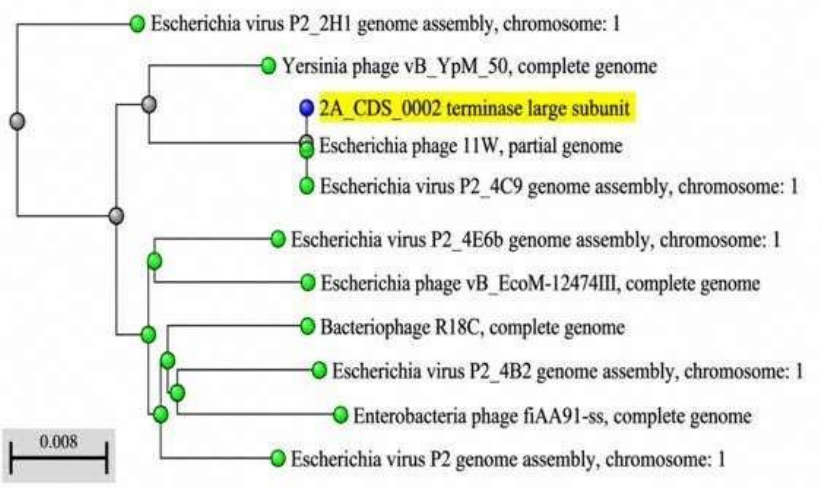


Figure 7. Terminase large-subunit (terL)-based phylogenetic tree of Phage-2 (2A) and related phage sequences.

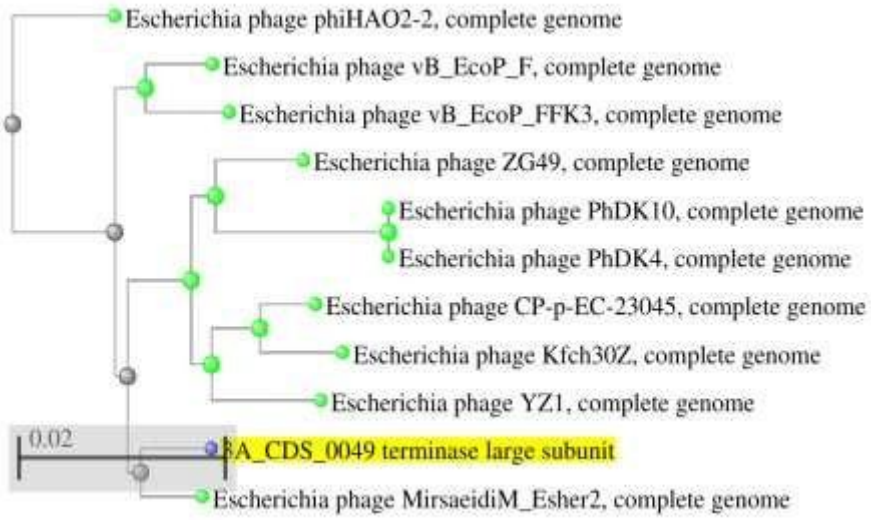


Figure 8. Terminase large-subunit (terL)-based phylogenetic tree of Phage-3 (3A) and related phage sequences.

3.10 Nucleotide composition

Table 9. Nucleotide composition of the assembled genomes.

Nucleotide	P1 count	P2 count	P3 count
A	13,253	7,583	10,414
T	14,740	7,373	9,524
G	10,357	8,549	10,553
C	11,993	7,670	9,208
GC (%)	44.395	51.384	49.777

Ambiguous N bases	0	269	0

The P2 assembly additionally contained other IUPAC ambiguity codes; its nucleotide composition should therefore be interpreted with these ambiguous bases in mind.

4. DISCUSSION

The present study provides a comparative genomic characterization of three bacteriophages active against MDR E. coli. The three assemblies differed substantially in genome size, ranging from 31,564 bp for P2 to 50,343 bp for P1, and also differed in GC content and functional annotation. Such variation is consistent with the genomic diversity expected among bacteriophages, even when their bacterial host range overlaps.

The assemblies were each represented by a single contig, with an N50 equal to the complete assembled length and an L50 of 1. P1 and P3 contained no ambiguous bases, whereas P2 contained 269 N bases and 85 reported gaps. The P2 assembly therefore warrants more cautious interpretation than the assemblies of P1 and P3, particularly for analyses that depend on complete nucleotide continuity.

MASH analysis identified very close database matches for all three study phages, with a distance of 0.0 and 1000/1000 matching hashes. MASH provides an efficient alignment-free measure of genome similarity and is useful for rapidly identifying related genomes. [5] These values indicate that P1, P2 and P3 represent new isolates of previously described phages rather than novel phage species; their significance lies in their recovery from local environmental water sources and their activity against contemporary clinical MDR E. coli isolates. The terL phylogeny placed P1 and P2 with vB_Eco_mar001J1 and 11W, respectively, consistent with the MASH findings, whereas the P3 terL clustered with Escherichia phage MirsaeidiM_Esher2, reflecting the limited set of BLASTN hits included in the single-gene tree. A closest-reference match should not be interpreted as proof that a study isolate is identical to the reference genome.

Genome annotation identified structural and functional categories expected in bacteriophage genomes, including head and packaging, tail, DNA/RNA metabolism and lysis-associated functions. The presence of holin, endolysin and spanin-related annotations in P1 and P2, and holin, amidase and spanin-related annotations in P3, provides a genomic basis for phage-mediated host-cell lysis. Annotation of phage genomes requires careful interpretation because a considerable proportion of phage proteins can remain functionally uncharacterized. [4]

No CARD antimicrobial-resistance-gene hits or VFDB virulence-factor hits were detected in the three genomes. These findings are supportive when considering phages for further study, but they should not be interpreted as an absolute safety guarantee. Database-dependent screening can only identify determinants represented in the databases and detectable by the applied annotation pipeline.

The integrase annotation in P2 is an important distinguishing feature. P1 and P3 had no integration/excision-associated annotation, whereas P2 contained one integrase. In addition, P2 clustered with P2-like phages of the genus Peduovirus, which are typically temperate. Together, these findings suggest that P2 may be capable of lysogeny; it should therefore not be labelled as strictly lytic, and it is a less suitable therapeutic candidate than P1 and P3 unless its lifestyle is clarified experimentally. This distinction is important because genomic predictions and experimentally observed lytic activity answer different questions.

The circular genome maps provide useful visual representations of gene organization and feature orientation, but they do not by themselves demonstrate physical genome circularity.

The principal strengths of the present analysis are the integration of raw-read quality assessment, single-contig assembly statistics, nucleotide composition, comparative genome analysis, standardized phage annotation, lysis-gene assessment, CARD/VFDB screening and terL-based phylogeny. The main limitations are the absence of sequencing-depth (coverage) data, the ambiguous bases and gaps in the P2 assembly, the dependence of resistance and virulence screening on reference databases, and the lack of experimental confirmation of phage lifestyle.

5. CONCLUSION

Whole-genome sequencing and comparative genomic analysis revealed distinct genomic characteristics among three bacteriophages active against MDR E. coli. P1, P2 and P3 formed single-contig assemblies of 50,343 bp, 31,564 bp and 39,699 bp, respectively, with different GC contents and functional genomic profiles. Comparative analysis identified close relationships with vB_Eco_mar001J1, phage 11W and phage K1F, respectively, and terL-based phylogeny supported the placements of P1 and P2. Lysis-associated functions were identified in all three phages, while no CARD antimicrobial-resistance or VFDB virulence-factor hits were detected. The integrase annotation and P2-like affiliation of P2 indicate that its lifestyle should be interpreted cautiously, whereas P1 and P3 are the more suitable candidates for further therapeutic evaluation. The study provides a genomic foundation for further characterization of these bacteriophages.

6. ETHICS STATEMENT

The present work forms part of the PhD thesis protocol "In vitro isolation of potential phages against multidrug resistant *Escherichia coli*: isolated from clinical samples", which was approved by the Institutional Ethics Committee of Santosh Deemed to be University, Ghaziabad (F. No. SU/R/2024/2403[3], dated 05 September 2024) and by the Institutional Ethics Committee of Heritage Institute of Medical Sciences, Varanasi (Ref. No. HIMS/IEC/317, dated 15 September 2025). This manuscript reports the analysis of whole-genome sequence data generated from purified bacteriophage preparations and did not involve the direct participation of human subjects.

7. FUNDING

This research received no external funding.

8. CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet. 2022;399(10325):629-655. doi:10.1016/S0140-6736(21)02724-0.

[2] Poirel L, Madec JY, Lupo A, Schink AK, Kieffer N, Nordmann P, et al. Antimicrobial resistance in Escherichia coli. Microbiol Spectr. 2018;6(4):ARBA-0026-2017. doi:10.1128/microbiolspec.ARBA-0026-2017.

[3] Kortright KE, Chan BK, Koff JL, Turner PE. Phage therapy: a renewed approach to combat antibiotic-resistant bacteria. Cell Host Microbe. 2019;25(2):219-232. doi:10.1016/j.chom.2019.01.014.

[4] Shen A, Millard A. Phage genome annotation: where to begin and end. Phage (New Rochelle). 2021;2(4):183-193. doi:10.1089/phage.2021.0015.

[5] Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:132. doi:10.1186/s13059-016-0997-x.

[6] Bouras G, Nepal R, Houtak G, Psaltis AJ, Wormald PJ, Vreugde S. Pharokka: a fast scalable bacteriophage annotation tool. Bioinformatics. 2023;39(1):btac776. doi:10.1093/bioinformatics/btac776.

[7] Biokart India Pvt. Ltd. Bacteriophage whole genome sequencing report: Project BKWGS2596; sample 1A. Bengaluru: Biokart Genomics Lab; 2026.

[8] Biokart India Pvt. Ltd. Bacteriophage whole genome sequencing report: Project BKWGS2596; sample 2A. Bengaluru: Biokart Genomics Lab; 2026.

[9] Biokart India Pvt. Ltd. Bacteriophage whole genome sequencing report: Project BKWGS2596; sample 3A. Bengaluru: Biokart Genomics Lab; 2026.

[10] Biokart India Pvt. Ltd. Whole genome sequencing QC report: Project BKWGS2596. Bengaluru: Biokart Genomics Lab; 2026.

[11] Andrews S. FastQC: a quality control tool for high throughput sequence data [Internet]. Cambridge: Babraham Bioinformatics; 2010 [cited 2026 Sep 27]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[12] Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047-3048. doi:10.1093/bioinformatics/btw354.

[13] Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13(6):e1005595. doi:10.1371/journal.pcbi.1005595.

[14] McNair K, Zhou C, Dinsdale EA, Souza B, Edwards RA. PHANOTATE: a novel approach to gene identification in phage genomes. Bioinformatics. 2019;35(22):4537-4542. doi:10.1093/bioinformatics/btz265.

[15] Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genom Bioinform. 2021;3(3):lqab067. doi:10.1093/nargab/lqab067.

[16] Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026-1028. doi:10.1038/nbt.3988.

[17] Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021;49(16):9077-9096. doi:10.1093/nar/gkab688.

[18] Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004;32(1):11-16. doi:10.1093/nar/gkh152.

[19] Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007;8:209. doi:10.1186/1471-2105-8-209.

[20] Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 2020;48(D1):D517-D525. doi:10.1093/nar/gkz935.

[21] Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic Acids Res. 2019;47(D1):D687-D692. doi:10.1093/nar/gky1080.

[22] Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Clokie M, et al. INfrastructure for a PHAge REference Database: identification of large-scale biases in the current collection of cultured phage genomes. Phage (New Rochelle). 2021;2(4):214-223. doi:10.1089/phage.2021.0007.

[23] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2.

[24] Krueger F. Trim Galore: a wrapper tool around Cutadapt and FastQC [Internet]. Cambridge: Babraham Bioinformatics [cited 2026 Oct 9]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

[25] Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes de novo assembler. Curr Protoc Bioinformatics. 2020;70(1):e102. doi:10.1002/cpbi.102.