Genome Sequences of Four Potentially Therapeutic Bacteriophages Infecting Shiga Toxin-Producing *Escherichia coli*

Carla Dias,[a] Carina Almeida,[b] 🔘 Małgorzata Łobocka,[c] 🔘 Hugo Oliveira[a]

[a]CEB–Centre of Biological Engineering, University of Minho, Braga, Portugal
[b]INIAV, IP–National Institute for Agrarian and Veterinary Research, Vairão, Vila do Conde, Portugal
[c]Department of Microbial Biochemistry, Institute of Biochemistry and Biophysics of the Polish Academy of Sciences, Warsaw, Poland

ABSTRACT Four phages infecting Shiga toxin-producing *Escherichia coli* (STEC) strains of different serotypes were isolated from wastewater samples. Their virion DNAs range from 51 to 170 kbp, are circularly permuted or have defined terminal repeats, and can encode 82 to 279 proteins. Despite their high similarity to other phages, only about 30% of their genes have a predicted function.

Shiga toxin-producing *Escherichia coli* (STEC) causes significant foodborne diseases in humans. Being generally nonpathogenic in ruminants, they use their gut as a natural reservoir. Transmission to humans occurs through the consumption of contaminated foods, such as raw or undercooked meat products, raw milk, and contaminated raw vegetables. Because fecal shedding is the major contamination source of carcasses, causing subsequent food recalls and human outbreaks, the role of the live animal in the production of a safe food product is critical. Here, we report the isolation of four broad STEC-infecting phages (vB_EcoM_Lutter [Lutter], vB_EcoM_Ozark [Ozark], vB_EcoM_Gotham [Gotham], and vB_EcoS_Chapo [Chapo]) isolated in Braga, Portugal.

Phages were isolated and produced as described previously (1). Briefly, sewage samples enriched with double-strength tryptic soy broth medium and STEC strains were grown overnight at 37°C with agitation. Filtered supernatants were spotted onto bacterial lawns, and collected phages were used for further purification.

Phage genomic DNA was extracted using phenol-chloroform-isoamyl alcohol extraction (2). Next, whole-genome libraries were constructed using a TruSeq Nano DNA library prep kit. The generated DNA fragments were multiplexed and sequenced in the same Illumina MiSeq run using 300-bp paired-end sequencing reads. The sequence reads were assembled in the Geneious Prime 2020 (Biomatters Ltd., New Zealand) *de novo* assembler (with medium-low sensitivity), yielding average coverages of 97× (61,819 reads), 20× (9,253 reads), 79× (31,782 reads), and 130× (19,306 reads) for Lutter, Ozark, Gotham, and Chapo, respectively. Quality control of the sequence reads was performed with FastQC v0.11.5 (3), while the assembly quality was verified with Geneious Prime (4). The assembled reads of Lutter, Ozark, and Chapo formed single contigs of overlapping ends with no regions of 2× increased coverage, as expected in the case of terminally redundant and circularly permuted sequences. Their starts were selected to align with the starts of the genomes of similar reference phages. The genomes were annotated using MyRAST (5), BLAST (6), tRNAscan-SE v2.0 (7), ARAGORN (8), PhagePromoter (9), and HHpred (10) (with default program parameters) and manually inspected. A summary of their basic characteristics is presented in Table 1.

Lutter was isolated using a STEC O104 strain. It is a myovirus with a 170,054-bp genome that can encode 279 putative proteins (only 120 with predicted function) and

TABLE 1 Morphology and overall features of isolated *Escherichia* phages

Phage name	Morphology (family)	Subfamily, genus	Genome size (bp)	Virion DNA	Packaging strategy	G+C content (%)	No. of CDS[a]	No. of tRNAs
vB_EcoM_Lutter	*Myoviridae* (*Myoviridae*)	*Tevenvirinae*, *Tequatrovirus*	170,054	Terminally redundant, circularly permuted	Headful packaging, preferred *pac* cuts between pos.[b] 97225 and 97248 of genomic sequence	35.4	279	8
vB_EcoM_Ozark	*Myoviridae* (*Myoviridae*)	*Tevenvirinae*, *Tequatrovirus*	167,600	Terminally redundant, circularly permuted	Headful packaging, preferred *pac* cuts between pos.[b] 94420 and 94443 of genomic sequence	39.5	268	10
vB_EcoM_Gotham	*Myoviridae* (*Myoviridae*)	*Vequintavirinae*, *Vequintavirus*	137,054	With 459-bp terminal repeats	Same specific start sequence for packaging of all virions	43.7	214	6
vB_EcoS_Chapo	*Siphoviridae* (*Drexlerviridae*)	*Tunavirinae*, *Tunavirus*	51,099	Terminally redundant, circularly permuted	Headful packaging, *pac* cut at pos.[b] 68/69 of genomic sequence	45.5	82	0

[a] CDS, coding DNA sequences.
[b] pos., position(s).

shares 90% overall nucleotide identity with the *Escherichia* phage teqhad (GenBank accession number MN895434). Ozark, isolated using a different STEC O29:H12 strain, is closely related to Lutter (97% overall nucleotide identity). They are both related to prototypical phage T4 and share the preferred 24-bp region of T4 DNA packaging. Gotham is a smaller myovirus with a 137,025-bp DNA molecule and 459-bp terminal repeats, sharing 90% overall nucleotide identity with several other *Escherichia* phages (e.g., vB_EcoM-ECP26, GenBank accession number MK883717). Chapo is a siphovirus isolated using the STEC O29:H12 strain and is related to phage T1. It has a 51,099-bp genome divided into oppositely transcribed halves and can encode 82 potential proteins (only 22 with predicted functions). The *pac* cut site of Chapo was localized between positions 68 and 69 of the genomic sequence pointed out by the identical ends in ~20% of these region reads. All the genomes have defined modules coding different functions. In particular, the lysis cassettes of the myoviruses contain putative holin and endolysin genes that are separated, with the exception of Gotham, where the holin gene was not identified. Siphovirus Chapo is predicted to encode a holin, an endolysin, and u-spanin canonical genes.

Data availability. The GenBank accession numbers are MT682713, MT682714, MT682715, and MT682716 for vB_EcoM_Ozark, vB_EcoM_Lutter, vB_EcoS_Chapo, and vB_EcoM_Gotham, respectively. The corresponding SRA data have been deposited in NCBI under BioProject accession number PRJNA646048.

REFERENCES

1. Oliveira H, Pinto G, Oliveira A, Oliveira C, Faustino MA, Briers Y, Domingues L, Azeredo J. 2016. Characterization and genome sequencing of a Citrobacter freundii phage CfP1 harboring a lysin active against multidrug-resistant isolates. Appl Microbiol Biotechnol 100:10543–10553. https://doi.org/10.1007/s00253-016-7858-0.
2. Sambrook JR. 2001. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, New York, NY.
3. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
4. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: an integrated and extendable

desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649. https://doi.org/10.1093/bioinformatics/bts199.
5. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: Rapid Annotations using Subsystems Technology. BMC Genomics 9:75. https://doi.org/10.1186/1471-2164-9-75.
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

7. Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964. https://doi.org/10.1093/nar/25.5.955.

8. Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32:11–16. https://doi.org/10.1093/nar/gkh152.

9. Sampaio M, Rocha M, Oliveira H, Dias O. 2019. Predicting promoters in phage genomes using PhagePromoter. Bioinformatics 35:5301–5302. https://doi.org/10.1093/bioinformatics/btz580.

10. Soding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244–W248. https://doi.org/10.1093/nar/gki408.