




Genome Annotation for Pinkcreek, a C1 Subcluster Mycobacteriophage from New Orleans, Louisiana

 Christine A. Byrum,^a Maximiliano F. Flota,^a Jackson A. Derrick,^a Steven Grant Dixon,^a Andre S. Gagliano,^a Halee S. Gibson,^a Paige A. Loose,^a Ashley Montero,^a Jessica L. Mugford,^a Emily E. Mullner,^a Briana C. Pope,^a Christopher A. Korey^a

^aDepartment of Biology, College of Charleston, Charleston, South Carolina, USA

ABSTRACT The mycobacteriophage Pinkcreek (C1 subcluster) was extracted from soil collected on the Dr. Norman C. Francis Parkway Bike Trail in New Orleans, Louisiana. It is a member of the family *Myoviridae* and infects *Mycobacterium smegmatis* mc²155. The Pinkcreek genome is 153,184 bp and contains 216 predicted protein-coding genes, 29 tRNAs, and 1 transfer-messenger RNA.

In a broad effort to better characterize viral diversity and evolution, the bacteriophage Pinkcreek was extracted from soil gathered on the Dr. Norman C. Francis Parkway Bike Trail in New Orleans, Louisiana (29.9619N, 90.1013W), during fall 2018 (Table 1). This project was sponsored by the Howard Hughes Medical Institute (HHMI) Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) program (1), and Pinkcreek was isolated by direct plating followed by two cycles of purification/amplification using 7H9 top agar containing *Mycobacterium smegmatis* mc²155 at 37°C, in accordance with the SEA-PHAGES Discovery Guide (2).

To sequence the genome, DNA was extracted from high-titer lysates using the Promega Wizard DNA cleanup system, and a sequencing library was prepared with the NEBNext Ultra II library prep kit (v3 reagents). Pittsburgh Bacteriophage Institute sequenced the DNA on an Illumina MiSeq system (MiSeq reagent kit v3) (3), and 382,828 single-end reads (150 bp) were obtained (coverage, 353×; average Phred score, 37.29). Raw reads were assembled *de novo* into a single contig using Newbler v2.9 (4), and editing and finishing were performed with Consed v29.0 (3, 5). Lack of read buildups (detected using PAUSE [<https://cpt.tamu.edu/computer-dresources/pause>]) indicated that the 153,184-bp genome (GC content, 64.6%) is circularly permuted. AceUtil (<http://phagesdb.org/AceUtil>) was utilized to check for sequence discrepancies and low-coverage sites. Further details were described by Russell (3).

Pinkcreek was annotated using the PECAAN workflow tool (6), with start sites determined using GeneMark v2.5 (7), GLIMMER v3.02 (8), and Starterator v1.1 (9); functional calls were made with HHpred (10), BLASTp v2.13.0+ (11), TOPCONS v2 (12), TMHMM2 (<https://services.healthtech.dtu.dk/service.php?TMHMM-2.0>), SOSUI v1.11 (13), and the NCBI Conserved Domain Database (CDD) (14), while tRNAs and transfer-messenger RNAs (tmRNAs) were identified using tRNAscan-SE v3.0 (15) and ARAGORN v1.2.38 (16). Parameters and databases used by PECAAN for the HHpred, BLASTp, and CDD searches are summarized at <https://seaphages.org/forums/topic/5398>. Other programs utilized default parameters. After annotation, data were transferred to DNA Master v5.22.2 (<https://phagesdb.org/DNAMaster>).

Based on nucleotide sequence similarities, phages are assigned to clusters sharing nucleotide sequence similarity of >50% (17) and/or gene content similarity of ≥35% (18). Pinkcreek, a C cluster/C1 subcluster member, is a lytic mycobacteriophage with *Myoviridae* morphology and a genome containing 216 predicted protein-coding genes (46 with assigned putative functions), 29 tRNAs, and 1 tmRNA. To compare Pinkcreek's

Editor Catherine Putonti, Loyola University Chicago

Copyright © 2022 Byrum et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Christine A. Byrum, byrumc@cofc.edu.

The authors declare no conflict of interest.

Received 10 June 2022

Accepted 2 August 2022

Published 18 August 2022

TABLE 1 Characteristics of the Pinkcreek bacteriophage

Parameter	Pinkcreek data
GenBank accession no.	MZ958745
SRA accession no.	SRX13720608
Collection site	New Orleans, Louisiana, USA
Collection site coordinates	29.9619N, 90.1013W
Isolation host	<i>Mycobacterium smegmatis</i> mc ² 155
Genome size (bp)	153,184
Coverage (×)	353
GC content (%)	64.7
No. of predicted protein-coding genes	216
No. of tRNAs	29
No. of tmRNAs	1
Morphotype	<i>Myoviridae</i>
Subcluster	C1
Predicted protein-coding genes (phams) unique to and conserved in all C1 subcluster members ^a	4, 5, 13, 23, 25, 29, 44, 50, 51, 54, 55 (helix-turn-helix DNA binding domain protein), 57, 59, 62, 68, 69, 82, 94, 98, 104, 107, 112 (acetyltransferase), 129, 132–134, 138, 139, 142, 206, 212–214, 231 (membrane protein), 234–237, 239, 240 (serine/threonine kinase), 241 (HNH endonuclease), 243 (PurA-like adenylosuccinate synthetase)

^a Based on data available in Phamerator on 1 June 2022 (19). Sequences with known predicted functions are indicated in parentheses.

genome to those of other actinobacteriophages, the program Phamerator (19) was used. Phamerator generates a map of the genome and, by selecting a single gene, users can access a pulldown menu listing all actinobacteriophage clusters (and cluster members) with the same pham (homologous protein-coding genes sharing $\geq 32.5\%$ identity). Pinkcreek's genome has 42 phams that are conserved in all C1 subcluster members ($n = 160$) but are absent in other actinobacteriophages (Table 1). The genome also contains an orpham (gp135), a tandem duplication (gp100/gp101), and a rare pham (gp12) encountered in only three other C1 subcluster members, namely, HyRo (GenBank accession number [KT281790](#)), Shifa (GenBank accession number [MT889395](#)), and Stubby (GenBank accession number [MK450423](#)). All genes are transcribed on the forward strand except gp39 to gp41, gp137 to gp138, and gp171 to gp172, and whole-genome BLASTn alignments (11) revealed that the C1 subcluster member Alice (GenBank accession number [JF704092](#)) (99.58% identity and 97% coverage) is most similar to Pinkcreek. Other similar C1 subcluster members ($\geq 99\%$ identity and $\geq 92\%$ coverage) include Blackbrain (GenBank accession number [MK878897](#)), Grungle (GenBank accession number [MN062707](#)), Koguma (GenBank accession number [MF919513](#)), LinStu (GenBank accession number [JN412592](#)), and Sauce (GenBank accession number [NC_054722](#)). Four tRNAs present in most of these phages are absent in Pinkcreek (would normally occur between base 92747 and base 92767).

Data availability. GenBank and SRA accession numbers are presented in Table 1.

ACKNOWLEDGMENTS

Thanks go to the HHMI SEA-PHAGES program for generous support, as well as the Department of Biology at the College of Charleston and SC INBRE (summer support for M.F.F. and publication costs were covered by a SC INBRE program grant from the National Institutes of Health National Institute of General Medical Sciences [grant P20GM103499-20]).

Special thanks go to Xavier University of Louisiana student Majesty Mason and his mentor Joseph Ross for initial isolation of Pinkcreek and its genome in the Introduction to Phage and Genomics course. We also thank Graham F. Hatfull, Deborah Jacobs-Sera, Daniel A. Russell, Welkin H. Pope, and Rebecca A. Garlena at the University of Pittsburgh for sequencing, quality control, and assembly of the genome and College of Charleston undergraduates Sharonda Cooper and Drew Pampu for their assistance with genome annotation. Finally, we thank Deborah Jacobs-Sera (University of Pittsburgh) and Richard Pollenz (University of South Florida) for reviewing the annotated genome and providing valuable feedback.

REFERENCES

- Jordan TC, Burnett SH, Carson S, Caruso SM, Clase K, DeJong RJ, Dennehy JJ, Denver DR, Dunbar D, Elgin SCR, Findley AM, Gissendanner CR, Golebiewska UP, Guild N, Hartzog GA, Grillo WH, Hollowell GP, Hughes LE, Johnson A, King RA, Lewis LO, Li W, Rosenzweig F, Rubin MR, Saha MS, Sandoz J, Shaffer CD, Taylor B, Temple L, Vazquez E, Ware VC, Barker LP, Bradley KW, Jacobs-Sera D, Pope WH, Russell DA, Cresawn SG, Lopatto D, Bailey CP, Hatfull GF. 2014. A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. *mBio* 5:e01051-13. <https://doi.org/10.1128/mBio.01051-13>.
- Poxleitner M, Pope W, Jacobs-Sera D, Sivanathan V, Hatfull G. 2018. Phage discovery guide. Howard Hughes Medical Institute, Chevy Chase, MD. <https://seaphagesphagediscoveryguide.helpdocsonline.com/home>.
- Russell DA. 2018. Sequencing, assembling, and finishing complete bacteriophage genomes. *Methods Mol Biol* 1681:109–125. https://doi.org/10.1007/978-1-4939-7343-9_9.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380. <https://doi.org/10.1038/nature03959>.
- Gordon D, Green P. 2013. Consed: a graphical editor for next-generation sequencing. *Bioinformatics* 29:2936–2937. <https://doi.org/10.1093/bioinformatics/btt515>.
- Rinehart CA, Gaffney BL, Smith JR, Wood JD. 2016. PECAAN: Phage Evidence Collection and Annotation Network. Western Kentucky University Bioinformatics and Information Science Center, Bowling Green, KY. https://seaphages.org/media/docs/PECAAN_User_Guide_Dec7_2016.pdf.
- Lukashin AV, Borodovsky M. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* 26:1107–1115. <https://doi.org/10.1093/nar/26.4.1107>.
- Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 27:4636–4641. <https://doi.org/10.1093/nar/27.23.4636>.
- Pacey M. 2016. Starterator guide. Pope W (ed). University of Pittsburgh, Pittsburgh, PA. https://seaphages.org/media/docs/Starterator_Guide_2016.pdf.
- Söding 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>.
- 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](https://doi.org/10.1016/S0022-2836(05)80360-2).
- Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. 2015. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res* 43:W401–W407. <https://doi.org/10.1093/nar/gkv485>.
- Hirokawa T, Boon-Chieng S, Mitaku S. 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 14:378–379. <https://doi.org/10.1093/bioinformatics/14.4.378>.
- Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. 2015. CDD: NCBI's Conserved Domain Database. *Nucleic Acids Res* 43:D222–D226. <https://doi.org/10.1093/nar/gku1221>.
- 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>.
- 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>.
- Hatfull GF, Jacobs-Sera D, Lawrence JG, Pope WH, Russell DA, Ko C-C, Weber RJ, Patel MC, Germane KL, Edgar RH, Hoyte NN, Bowman CA, Tantoco AT, Paladin EC, Myers MS, Smith AL, Grace MS, Pham TT, O'Brien MB, Vogelsberger AM, Hryckowian AJ, Wynalek JL, Donis-Keller H, Bogel MW, Peebles CL, Cresawn SG, Hendrix RW. 2010. Comparative genomic analysis of 60 mycobacteriophage genomes: genome clustering, gene acquisition and gene size. *J Mol Biol* 397:119–143. <https://doi.org/10.1016/j.jmb.2010.01.011>.
- Mavrich TN, Hatfull GF. 2017. Bacteriophage evolution differs by host, lifestyle and genome. *Nat Microbiol* 2:17112. <https://doi.org/10.1038/nmicrobiol.2017.112>.
- Cresawn SG, Bogel M, Day N, Jacobs-Sera D, Hendrix RW, Hatfull GF. 2011. Phamerator: a bioinformatic tool for comparative bacteriophage genomics. *BMC Bioinformatics* 12:395. <https://doi.org/10.1186/1471-2105-12-395>.