




# Assembly and Annotation of *Escherichia coli* Bacteriophage U115

William An,<sup>a</sup> Camilla Emsbo,<sup>a</sup> Emma Frey,<sup>a</sup> Vicky Hu,<sup>a</sup> Ashley Jones,<sup>a</sup> Nipa Latif,<sup>a</sup> Makayla Perrilli,<sup>a</sup> Krystalia Reillo,<sup>a</sup> Joshua C. Schwarz,<sup>a</sup> Sydney Strasner,<sup>a</sup> Austin Theroux,<sup>a</sup> Luz D. Vargas,<sup>a</sup> Paul E. Turner,<sup>a,b,c</sup>  Alita R. Burmeister<sup>a,b</sup>

<sup>a</sup>Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut, USA

<sup>b</sup>BEACON Center for the Study of Evolution in Action, East Lansing, Michigan, USA

<sup>c</sup>Program in Microbiology, Yale School of Medicine, New Haven, Connecticut, USA

William An, Camilla Emsbo, Emma Frey, Vicky Hu, Ashley Jones, Nipa Latif, Makayla Perrilli, Krystalia Reillo, Joshua C. Schwarz, Sydney Strasner, Austin Theroux, and Luz D. Vargas contributed equally to this work. Order of these authors is given alphabetically by last name.

**ABSTRACT** We present the annotated genome sequence of *Escherichia coli* bacteriophage U115, a T4-like bacteriophage. Phage U115 has a genome length of 166,986 bp and has 286 predicted genes.

Previously described *Escherichia coli* phage U115 is of interest to the study of evolutionary trade-ups (1). The phage relies on the Tsx receptor, an outer membrane protein that also imports the antibiotic albicidin (1, 2). Consequently, an evolutionary trade-up can occur when phage resistance evolves through changes to Tsx that also block albicidin entry (1, 3). Phage U115 was isolated from wastewater influent in New Haven, CT, and characterized previously (1). A sample of phage U115 used in the current study was provided by Ben Chan (Yale University). We propagated the phage in Luria-Bertani broth at 37°C on *E. coli* K-12 strain BW25113.

DNA was isolated using a phage DNA isolation kit (Norgen Biotek). The sequencing library was prepared using the Illumina Nextera sequencing kit and sequenced on a Nextseq 2000 machine with a 300-cycle cartridge to give 150-bp reads. Genome assembly and annotation were conducted through the Center for Phage Technology (CPT) instances of Galaxy (4) and Web Apollo (5). Default parameters were used for all software unless otherwise specified. Sequences were rarified to a target coverage of 250× to improve assembly (6) using FASTQ Subset (7, 8). Sequence quality was assessed with FastQC v.0.72+galaxy1 (9). Low-quality sequence ends were trimmed with Trim Sequences v.1.0.2+galaxy0 (10) to reach a mean quality score above 30 across all bases and a 10th percentile quality score above 25 across all bases; this process involved trimming the first 18 bases and the last 1 base of each sequence. We also qualitatively confirmed that the per base sequence content of the trimmed sequences was consistent across the trimmed read lengths. The trimmed sequences were assembled using SPAdes v.3.12.0 (11) resulting in a circular contig of 166,986 bp and 119× coverage containing 55 bp of terminal overlapping sequence, which was removed manually before further analysis.

Using NCBI BLASTn (12), phage U115 was determined to be a teqatrovirus with a 94.63% identity and an 89% query coverage to phage T4 (GenBank accession number [MT984581](https://www.ncbi.nlm.nih.gov/nuclseq/MT984581)). The genome was reopened to be syntenic with phage T4.

The assembled sequence was run through the Galaxy phage annotation pipeline (PAP) structural workflow v.2021.02 (7) and imported into Apollo for structural annotation. Gene locations were predicted using GLIMMER3 v.0.2 (13), MetaGeneAnnotator v.1.0.0 (14), and Sixpack v.5.0.0+galaxy2 (15). tRNA gene calls were made with tRNAscan-SE v.2.0.5 (16) and ARAGORN v.0.6 (17). Finalized gene calls were confirmed

**Editor** John J. Dennehy, Queens College CUNY

**Copyright** © 2022 An 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 Alita R. Burmeister, [alita.burmeister@yale.edu](mailto:alita.burmeister@yale.edu).

**Received** 22 September 2021

**Accepted** 20 January 2022

**Published** 17 February 2022

by manual assessment of Shine-Dalgarno sequences, start and stop sequences, and distance between genes (18).

Functional annotation was initiated using the CPT PAP functional workflow v.2021.01 (7). Putative gene functions were assigned by manual assessment of BLASTp (19) results to the curated databases Canonical Phages, Swiss-Prot, and nonredundant (NR; NR phages only) along with InterProScan v.5.48-83.0 (20). Annotations were further verified separately using NCBI BLASTx (12) and InterPro (21).

Of the 286 predicted genes of phage U115, 144 were annotated as hypothetical genes, 131 were annotated with putative functions, and 11 were annotated as tRNA genes. EDGE bioinformatics (22) determined that the GC content of this genome was 35.42%, and there were no hits for virulence factors and deleterious genetic markers. PHACTS (23) predicted phage U115 to be “confidently lytic.”

This study uniquely represents a collaborative effort by 12 undergraduate researchers enrolled at 4 institutions, as follows: The University of New Haven (A.J., and K.R., and S.S.), Quinnipiac University (J.C.S., M.P., N.L., and E.F.), Southern Connecticut State University (L.D.V. and V.H.), and Yale University (A.T., W.A., and C.E.). All of the structural and functional annotations were completed during a 5-week, fully online summer research experience led by A.R.B.

**Data availability.** The annotated genome of phage U115 has been added to NCBI GenBank under accession number [MZ753803](#). The BioProject accession number is [PRJNA753771](#), and the Sequence Read Archive (SRA) number is [SRR15420633](#).

## ACKNOWLEDGMENTS

This research was funded by Howard Hughes Medical Institute Campus grant number 52008128 to P.E.T., by NIH National Institute of Allergy and Infectious Diseases grant number R21AI144345 to A.R.B. and P.E.T.

We thank Curtis Ross and Anthony Criscione at the CPT for their support and guidance, Carli Roush for her helpful insight, Katie Kortright for information on phage U115, and Ben Chan for a copy of phage U115.

The authors declare a conflict of interest. P.E.T. is a co-founder of Felix Biotechnology Inc., and declares a financial interest in this company that seeks to commercially develop phages for use as therapeutics. P.E.T. has two provisional patent applications related to phage therapy. A.R.B. has two provisional patent applications related to phage therapy.

## REFERENCES

- Kortright KE, Doss-Gollin S, Chan BK, Turner PE. 2021. Evolution of bacterial cross-resistance to lytic phages and albicidin antibiotic. *Front Microbiol* 12:658374. <https://doi.org/10.3389/fmicb.2021.658374>.
- Kortright KE, Chan BK, Turner PE. 2020. High-throughput discovery of phage receptors using transposon insertion sequencing of bacteria. *Proc Natl Acad Sci U S A* 117:18670–18679. <https://doi.org/10.1073/pnas.2001888117>.
- Burmeister AR, Turner PE. 2020. Trading-off and trading-up in the world of bacteria-phage evolution. *Curr Biol* 30:R1120–R1124. <https://doi.org/10.1016/j.cub.2020.07.036>.
- Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltmann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 46:W537–W544. <https://doi.org/10.1093/nar/gky379>.
- Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elisk CG, Lewis SE. 2013. Web Apollo: a web-based genomic annotation editing platform. *Genome Biol* 14:R93. <https://doi.org/10.1186/gb-2013-14-8-r93>.
- Phillipson CW, Voegtly LJ, Lueder MR, Long KA, Rice GK, Frey KG, Biswas B, Cer RZ, Hamilton T, Bishop-Lilly KA. 2018. Characterizing phage genomes for therapeutic applications. *Viruses* 10:188. <https://doi.org/10.3390/v10040188>.
- Ramsey J, Rasche H, Maughmer C, Criscione A, Mijalis E, Liu M, Hu JC, Young R, Gill JJ. 2020. Galaxy and Apollo as a biologist-friendly interface for high-quality cooperative phage genome annotation. *PLoS Comput Biol* 16:e1008214. <https://doi.org/10.1371/journal.pcbi.1008214>.
- E. Mijalis HR. 2013–2017. CPT Galaxy Tools. <https://github.com/tamu-cpt/galaxy-tools/>.
- Andrews S. FastQC: a quality control tool for high throughput sequence data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Gordon A. 2010. FASTQ/A short-reads pre-processing tools. [http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/).
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. <https://doi.org/10.1089/cmb.2012.0021>.
- NCBI Resource Coordinators. 2018. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 46:D8–D13. <https://doi.org/10.1093/nar/gkx1095>.
- 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>.
- Noguchi H, Taniguchi T, Itoh T. 2008. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res* 15:387–396. <https://doi.org/10.1093/dnares/dsn027>.
- Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P,

- Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* 47:W636–W641. <https://doi.org/10.1093/nar/gkz268>.
16. Lowe TM, Chan PP. 2016. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res* 44:W54–W57. <https://doi.org/10.1093/nar/gkw413>.
  17. 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>.
  18. Schlub TE, Holmes EC. 2020. Properties and abundance of overlapping genes in viruses. *Virus Evol* 6:veaa009. <https://doi.org/10.1093/ve/veaa009>.
  19. Cock PJA, Chilton JM, Grüning B, Johnson JE, Soranzo N. 2015. NCBI BLAST+ integrated into Galaxy. *Gigascience* 4:39. <https://doi.org/10.1186/s13742-015-0080-7>.
  20. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30: 1236–1240. <https://doi.org/10.1093/bioinformatics/btu031>.
  21. Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, Richardson L, Salazar GA, Williams L, Bork P, Bridge A, Gough J, Haft DH, Letunic I, Marchler-Bauer A, Mi H, Natale DA, Necci M, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A, Finn RD. 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* 49:D344–D354. <https://doi.org/10.1093/nar/gkaa977>.
  22. Li P-E, Lo C-C, Anderson JJ, Davenport KW, Bishop-Lilly KA, Xu Y, Ahmed S, Feng S, Mokashi VP, Chain PSG. 2017. Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Res* 45:67–80. <https://doi.org/10.1093/nar/gkw1027>.
  23. McNair K, Bailey BA, Edwards RA. 2012. PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics* 28:614–618. <https://doi.org/10.1093/bioinformatics/bts014>.