Poster Number: 218
February 25, 2026

From fragments to full circles: A scalable strategy for complete phage genome assembly

ATCC®
Credible leads to Incredible®

Joseph R. Petrone, PhD; Emily A. White, MS; Nikhita Puthuveetil, MS; Noah Wax, MS; Corina Tabron, MS; Stephen King, MS; Ana Fernandes, BS; John Bagnoli, BS; Jonathan L. Jacobs, PhD
ATCC, Manassas, VA 20110

Introduction

In this study, we introduce a targeted approach for improving phage genome assembly by leveraging high-quality, authenticated bacterial reference genomes available through the ATCC® Genome Portal (AGP). Sequenced reads from phage extractions were aligned to these curated reference genomes to perform precise in silico de-hosting, enabling the removal of host-derived bacterial sequences prior to assembly. Because short-read assemblies alone often result in fragmented or incomplete phage genomes, we supplemented this approach with Oxford Nanopore Technologies (ONT) long-read sequencing of phage-enriched material. The combined data were assembled using Autocycler, a pipeline optimized for circular genome detection and resolution.

ATCC Genome Portal

As of January 2026, 53 phages have been assembled, curated, and deposited in the AGP, representing 17 bacterial hosts and across three Baltimore classes. While this collection of phage genomes spans a broad range of host specificity, notable intra-strain infectivity is observed in *Escherichia coli* (ATCC® 11303™), from which 12 of its characterized phages are deposited on the AGP.

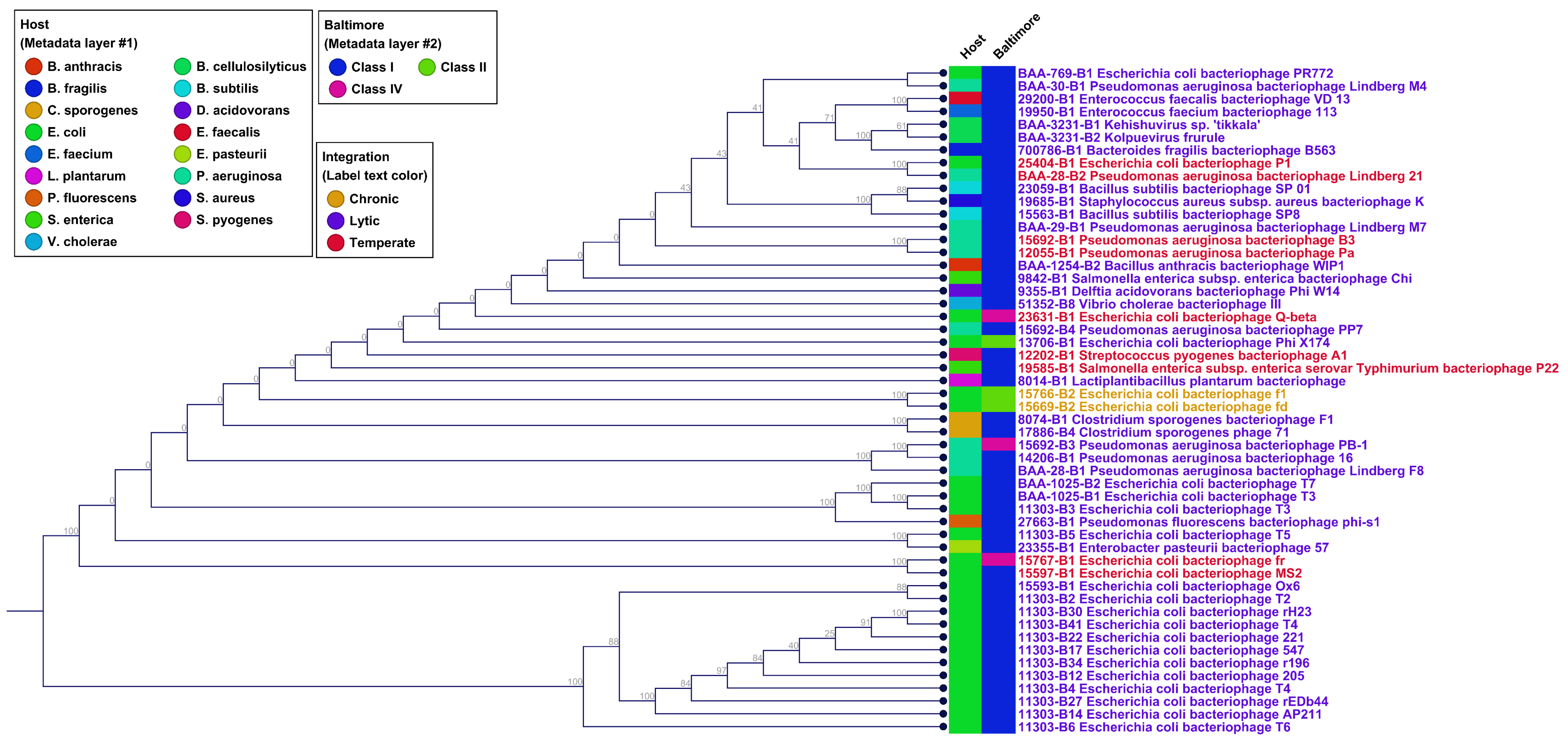

Figure 1: Phylogenetic tree of the 53 phages currently deposited on the AGP. The whole-genome alignment was created in CLC Workbench 25 using the Neighbor Joining algorithm with the Jukes-Cantor distance measurement. The host organism of the phage can be seen colored by the first metadata column with the Baltimore class colored by the second metadata layer. Phage labels are colored by integration lifestyle.

Difficult Phages To Assemble

Table 1: Difficult-to-assemble phages and their solutions prior to publication on the AGP.

Phage Name	Genome Characteristics	Difficult Features	Length Assembly / Ref*	Solution
Pseudomonas putida bacteriophage gh-1 (ATCC® 12633-B1™)	▪ dsDNA ▪ Lytic	▪ Terminal repeats (216 bp)	35,177 bp / 37,359 bp*	Re-extract; prep DNA for high molecular weight ONT
Escherichia coli bacteriophage lambda (λ) (ATCC® 23724-B2™)	▪ dsDNA ▪ Temperate ▪ 12-bp cohesive ends (circular)	>15 NCBI E. coli references contaminated with λ phage	48,502 bp / 48,507 bp*	De-host with AGP Escherichia coli strain C600 genome
Listeria monocytogenes phage 243 (ATCC® 23074-B1™)	▪ Siphoviridae-type dsDNA ▪ Temperate	▪ Circularly permuted ▪ Terminally redundant ▪ Integrate at tRNA loci	38.9 kb / Unknown* (A500 ~35–43 kb)	▪ Mask prophage-like regions in host references ▪ Do not force circularization

(*) indicates closest reference available on NCBI Genomes.

Methods

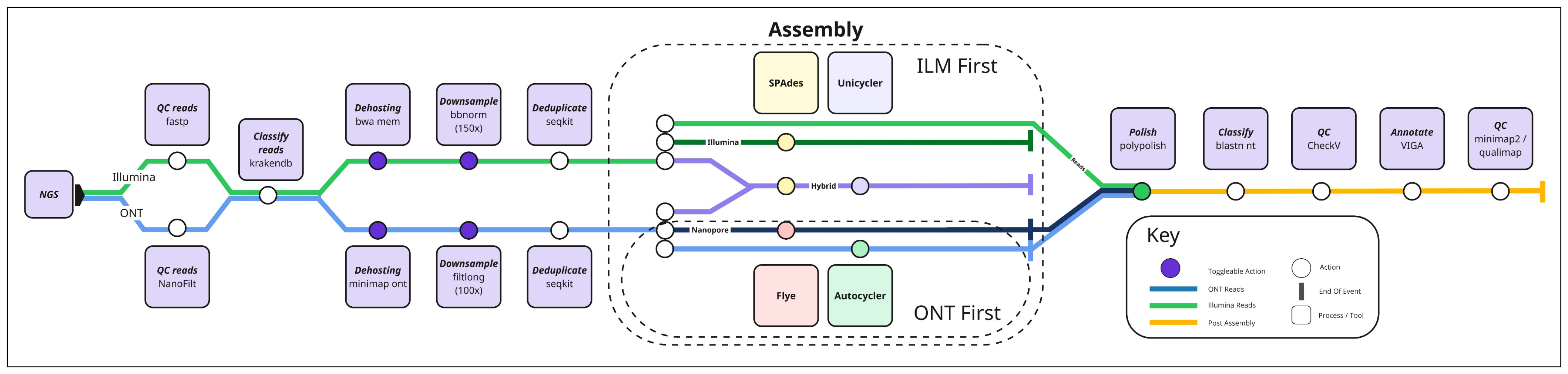

Figure 2: Graphical abstract of the bacteriophage assembly pipeline utilized for publication to the AGP. Strategies to produce contiguous and accurate assemblies include dehosting using host reference genome, downsampling, and varying assembler style. Diagram created in Miro.com.

Assembly Benchmarking

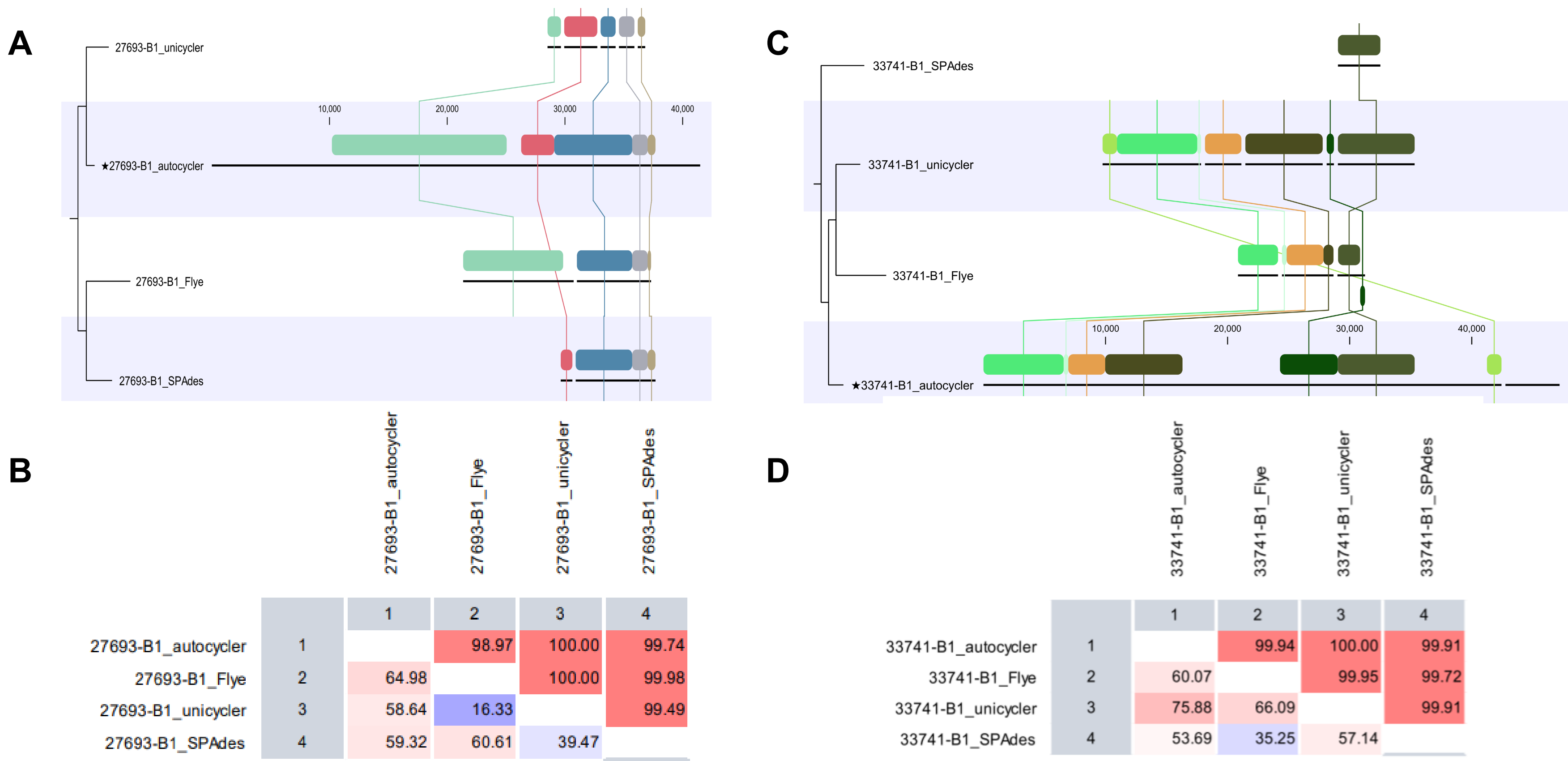

Figure 3: Assembly comparisons. Whole-genome alignments (WGAs) of various assembly attempts for bacteriophages in the ATCC® catalog currently under curation, demonstrating that Autocycler produces more contiguous phage genomes. WGAs, ANI, and images were done in CLC Genomics Workbench 26. (A&C) WGAs and LCBs are shown, and (B&D) ANI figures display percent ANI above the diagonal and aligned percentage below the diagonal. Benchmarking was performed with (A&B) ATCC® 27693-B1™ and (C & D) ATCC® 33741-B1™.

Table 2: Results summary. Assemblies generated with our standard viral pipeline (SPAdes), Flye, or Unicycler are much smaller in length than expected based on NCBI references, while assemblies generated with Autocycler more closely match expected phage lengths. Contig sizes are consistent across assemblers.

		Assembly Length (bp) / Contigs			
ATCC® Catalog Number	Expected Length (NCBI)	SPAdes	Flye	Unicycler	Autocycler
33741-B1™	42,431	3,471 / 1	9,753 / 3	24,264 / 5	46,870 / 2
27693-B1™	41,690	7,770 / 2	15,682 / 2	7,173 / 5	41,492 / 1
23724-B2™	40,931	No viral contigs found	40,911 / 1	No viral contigs found	48,502 / 1

Conclusions

- Combining curated reference-guided de-hosting with long-read sequencing and hybrid polishing produces significantly improved assemblies as compared to traditional short-read-only shotgun approaches.
- This work underscores the importance of curated reference genomes and tailored assembly strategies in phage genomics.

References

1. Wick RR, et al. Autocycler: long-read consensus assembly for bacterial genomes. Bioinformatics 41(9): btaf474, 2025. PubMed: 40875535.
2. Bankevich A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19(5): 455-477, 2012. PubMed: 22506599.
3. Kolmogorov M, et al. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37(5): 540-546, 2019. PubMed: 30936562.

ATCC 10801 University Boulevard, Manassas, Virginia 20110-2209 **Phone:** 800.638.6597 **Email:** sales@atcc.org **Web:** www.atcc.org