MPMI Vol. 35, No. 6, 2022, pp. 500–504, https://doi.org/10.1094/MPMI-01-22-0008-A

Complete Genome Sequences of Four Strains of *Erwinia tracheiphila*: A Resource for Studying a Bacterial Plant Pathogen with a Highly Complex Genome

Breah LaSarre, Olakunle I. Olawole, Ashley A. Paulsen, Larry J. Halverson, Mark L. Gleason, and Gwyn A. Beattie†

Department of Plant Pathology and Microbiology, Iowa State University, Ames, IA 50011-1101, U.S.A.

Funding
This project was funded by an Iowa State University Presidential Fellowship to O. I. Olawole, the College of Agriculture and Life Sciences of Iowa State University, and the United States Department of Agriculture National Institute of Food and Agriculture grant number 2021-67019-34833. A portion of this material is based upon work supported by the National Science Foundation under grant number DGE-1545453.

Keywords
bacterial wilt, complete genome, cucurbit wilt, *Erwinia tracheiphila*, genome assembly, insertion sequence, Nanopore, PacBio, prophage, xylem pathogen

Genome Announcement

Erwinia tracheiphila, the causative agent of bacterial wilt of cucurbits, has a highly complex genome harboring an abundance of repetitive elements and prophage. Here, we present the closed genome sequences of *E. tracheiphila* strains BHKY, BuffGH, MDCuke, and SCR3, which belong to two phylogenetic clades that differ in host-specific virulence. These are the first complete genome assemblies of this plant pathogen.

Erwinia tracheiphila, a gram-negative, xylem-limited, obligately insect-vectored plant pathogen, causes wilt of cucurbits belonging to the genera *Cucurbita* (squash and pumpkin) and *Cucumis* (melon and cucumber) (Saalau Rojas et al. 2015). *E. tracheiphila* poses a serious threat to commercial cucurbit production, with current disease management relying primarily on controlling the insect vector (Cavanagh et al. 2009; Sánchez et al. 2015; Weber 2018). Previous studies found that *E. tracheiphila* strains predominantly cluster into two clades that differ in host specificity: clade 'C-1' (*Et-C1*), which causes wilt in cucumber, melon, and squash, and clade 'melo' (*Et-melo*), which causes wilt in cucumber and melon but not in squash (Saalau Rojas et al. 2013; Shapiro et al. 2018b; Vrisman et al. 2016). Previous work also revealed that the *E. tracheiphila* genome is highly complex, containing an abundance of repetitive elements (e.g., insertion sequences) and prophage (Shapiro et al. 2016, 2018b), which has heretofore impeded the assembly of a closed genome for this pathogen (Shapiro et al. 2015, 2018a). Here, we leveraged Oxford Nanopore Technology (ONT) ultra-long read sequencing in combination with PacBio long-read or Illumina short-read sequencing to assemble complete, closed genomes for two *Et-C1* strains (BHKY and BuffGH) and two *Et-melo* strains (MDCuke and SCR3). These complete genome sequences will improve comparative genomic studies of *E. tracheiphila* and hold value for ongoing disease control efforts.

The four *E. tracheiphila* strains sequenced in this study were isolated from symptomatic plants in 2009 or 2010 at various locations in the United States (Table 1) (Saalau Rojas and Gleason 2012; Saalau Rojas et al. 2013). All genomic DNA was isolated from single-colony cultures grown in King's B broth (King et al. 1954) at 30°C with orbital shaking. Culturing and DNA isolation to prepare samples for sequencing were performed independently for each of

†Corresponding author: G. A. Beattie; gbeattie@iastate.edu

The author(s) declare no conflict of interest.

Accepted for publication 2 March 2022.

Table 1. Isolate information, sequencing and assembly metrics, and accession numbers of four complete *Erwinia tracheiphila* genome sequences[a]

Characteristic	Strains			
	BHKY	BuffGH	MDCuke	SCR3
Isolation source, location, year	Squash (*Cucurbita moschata*), Kentucky, 2010	Texas gourd (*Cucurbita pepo* subsp. *texana*), Pennsylvania, 2009	Cucumber (*Cucumis sativus*), Maryland, 2010	Muskmelon (*Cucumis melo*), Iowa, 2009
E. tracheiphila clade	C1	C1	Melo	Melo
Sequencing type (depth)	ONT (71X), PacBio (237X)	ONT (71X), Illumina (66X)	ONT (36X), Illumina (37X)	ONT (38X), PacBio (164X)
No. filtered ONT reads (N_{50} value [bp])	33,287 (18,370)	36,203 (16,335)	18,026 (17,892)	19,161 (18,970)
No. filtered PacBio reads (N_{50} value [bp])	185,114 (11,517)	N/A	N/A	127,264 (11,916)
No. filtered Illumina reads	N/A	3,426,896	2,001,694	N/A
Total genome size (bp)	4,958,521	4,978,853	5,054,064	4,974,774
Chromosome size (bp)	4,920,713	4,938,018	4,874,086	4,815,509
No. of plasmids	1	1	5	5
Total plasmid size (bp)[b]	37,808	40,835	179,978	159,265
G+C content (%)	50.51	50.57	50.49	50.49
No. of rRNAs (5S, 16S, 23S)	7, 6, 6	7, 6, 6	7, 6, 6	7, 6, 6
No. of tRNAs	65	65	65	65
Total no. of CDS[c]	5,043	5,096	5,283	5,168
No. of pseudogenes	847	881	888	873
BUSCO completeness score (%)	99.4	99.4	99.6	99.6
BioSample accession no.	SAMN24011481	SAMN24011482	SAMN24011483	SAMN24011484
Assembly accession no.	CP089932, CP089933	CP089940, CP089941	CP089942, CP089943, CP089944, CP089945, CP089946, CP089947	CP089934, CP089935, CP089936, CP089937, CP089938, CP089939
SRA accession no.[d]	SRR17231631 (O) SRR17231625 (P)	SRR17231630 (O) SRR17231627 (I)	SRR17231629 (O) SRR17231626 (I)	SRR17231628 (O) SRR17231624 (P)

[a] ONT = Oxford Nanopore Technologies; N/A = not applicable.
[b] BHKY has plasmid pETR004-b (37,808 bp); BuffGH has plasmid pETR004-c (40,835 bp); MDCuke has five plasmids: pETR001-a (71,258 bp), pETR002-a (39,625 bp), pETR003-a (31,842 bp), pETR004-a (30,315 bp), and pETR005-a (6,938 bp); and SCR3 has five plasmids: pETR001-b (68,396 bp), pETR002-b (39,626 bp), pETR006-a (35,983 bp), pETR007-a (8,238 bp), and pETR005-b (7,022 bp).
[c] CDS = coding DNA sequences.
[d] O = ONT, P = PacBio, and I = Illumina.

the three sequencing platforms. All DNA library preparation and sequencing was performed by or in collaboration with the Iowa State University DNA facility.

For ONT sequencing, high–molecular weight DNA was extracted from 25-ml cultures (optical density at 600 nm [OD_{600}] of 0.75 to 0.85) using a modified Sambrook and Russell phenol-chloroform-based protocol (available online), followed by size-selective precipitation of fragments >30 kb using polyethylene glycol (Lis and Schleif 1975), and final purification using AMPure XP beads (Beckmann Coulter). DNA libraries were prepared using an SQK-LSK109 ligation sequencing kit (ONT) and were sequenced on a MinION R9.4 flow cell for 72 h, with data acquisition using MinKNOW v21.05.12 (ONT), demultiplexing and base calling of raw data using Guppy v5.0.12 (ONT) using the high-accuracy model, and a read pass threshold of min_qscore = 9. Passed ONT reads were adapter-trimmed using Porechop v0.2.4, trimmed reads were then filtered for length (≥1,000 bp), using NanoFilt v2.8.0 (De Coster et al. 2018), and the filtered reads were assessed for quality using Nano-Plot v1.38.1 and NanoStat v1.5.0 (De Coster et al. 2018).

For PacBio sequencing of strains BHKY and SCR3 and Illumina sequencing of strains BuffGH and MDCuke, DNA was extracted from suspensions of late–log phase cultures that were adjusted to an OD_{600} of 1.0. Genomic DNA was purified from 1 ml of the adjusted suspension using the DNeasy blood and tissue kit (Qiagen), following manufacturer instructions and including steps for Proteinase K and RNase treatments, as recommended for PacBio genomic DNA extractions (Mayjonade et al. 2016). PacBio sequencing libraries were

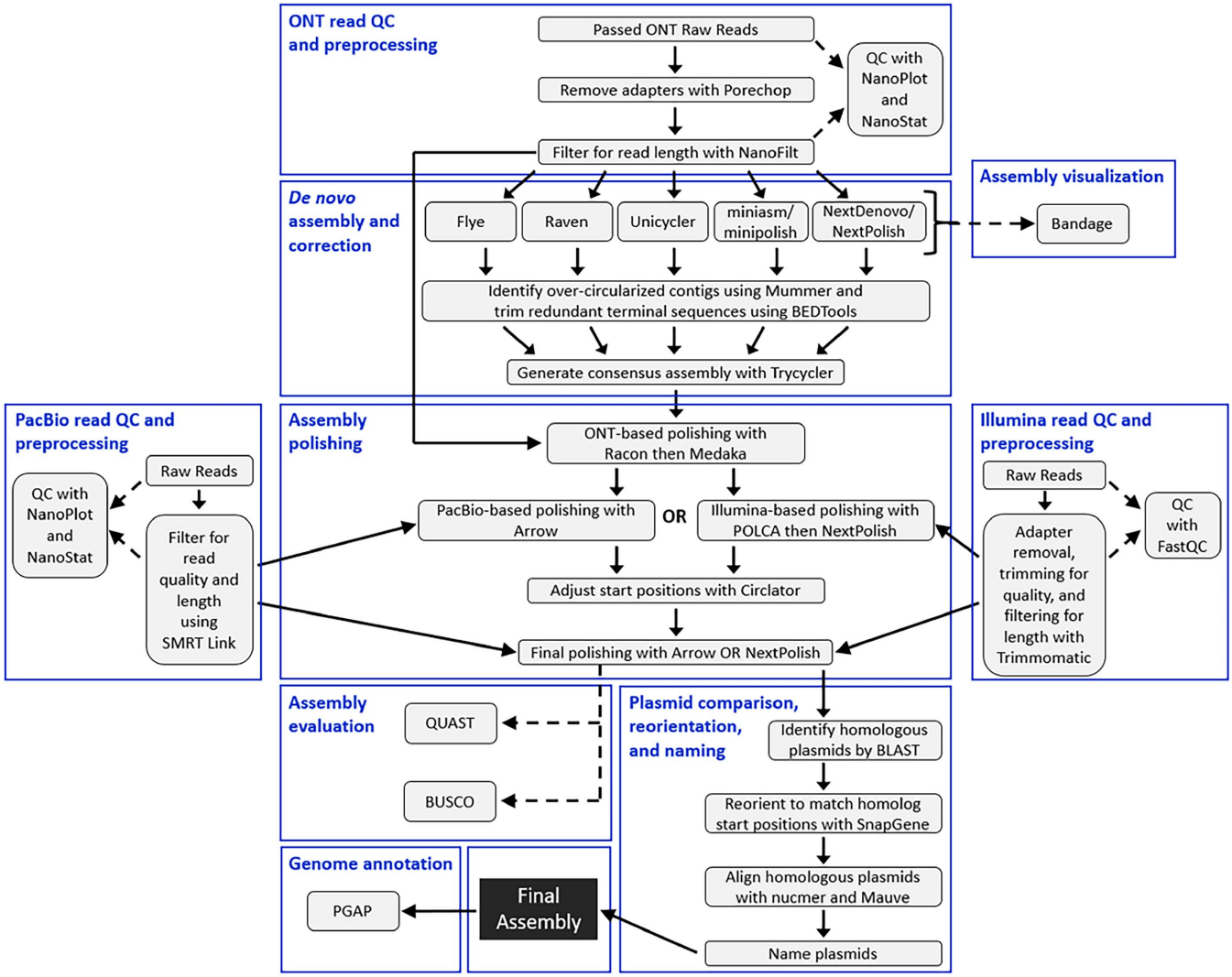

Fig. 1. Flow chart illustrating a pipeline for complex bacterial genome assembly. This pipeline utilizes a consensus (generated using Trycycler) of the output of multiple de novo assemblers (Flye, Raven, Unicycler, miniasm/minipolish, NextDenovo/NextPolish), followed by iterative polishing using data from multiple sequencing platforms. The assembly was evaluated with QUAST and BUSCO after each round of polishing. Steps assessing read or assembly quality are indicated with dashed arrows. ONT = Oxford Nanopore Technologies, QC = quality control.

prepared using the SMRTbell template prep kit (Pacific Biosciences), followed by 20-kb size selection using the BluePippin size-selection system (Sae Science Inc.). The libraries were sequenced using P6-C4 chemistry and default parameters on a PacBio RSII instrument, with one single-molecule real-time (SMRT) cell per strain. PacBio reads were filtered for length (≥1,000 bp) and quality (≥0.80) using SMRT Link v7.0.1 (Pacific Biosciences), and read quality was assessed using NanoPlot and NanoStat. Illumina sequencing libraries were prepared using the Nextera DNA flex library prep kit (Illumina) and were sequenced using the Illumina HiSeq3000 platform (2× 100-bp read length) with the NEBNext Ultra II FS kit (New England Biolabs), with data acquisition using HiSeq control software HD 3.4.0.38 (Illumina) and demultiplexing using bcl2fastq v2.20.0.422 (Illumina). Illumina reads were adapter and quality trimmed using Trimmomatic v0.39 (keepBothReads LEADING:3 TRAILING:3 SLIDINGWINDOW:4:25 MINLEN:40) (Bolger et al. 2014), and the quality of the trimmed reads was assessed using FastQC v0.11.7.

The *E. tracheiphila* genomes were assembled using a multiassembly consensus plus polishing approach (Fig. 1). All software programs were run using default parameters unless otherwise specified. For each strain, the filtered ONT reads were used to generate five independent de novo assemblies using the following programs: Flye v2.9 (Kolmogorov et al. 2019), Raven v1.6.1 (Vaser and Šikić 2021), Unicycler v0.4.9 (long-read only) (Wick et al.

2017), Miniasm/Minipolish (v0.3/v0.1.3) (Wick and Holt 2021), and NextDenovo/NextPolish (v2.5.0/v1.4.0) (Wick and Holt 2021). Assembly graphs were visualized using Bandage v0.8.1 (Wick et al. 2015). For each assembly, over-circularized contigs (i.e., contigs with near-identical terminal overlaps) were identified by self-versus-self alignment with Mummer v3.23 (nucmer –maxmatch –nosimplify) (Delcher et al. 2002); terminal overlaps were subsequently resolved using the BEDtools v2.27.1 tool 'getfasta' (Quinlan and Hall 2010), resulting in a trimmed version of each assembly. The five trimmed assemblies for a given strain were then merged into a single consensus assembly using Trycycler v0.4.1 (–max_indel_size 500) (Wick et al. 2021). Each consensus assembly was initially polished using filtered ONT reads with two iterations with Racon v1.6.1 (-m 8 -x 6 -g -8 -w 500) (Vaser et al. 2017) and then one iteration (1×) with Medaka v1.4.1 (-m r941_min_hac_g507) (ONT). The ONT-polished assemblies were then further polished using either filtered PacBio reads with Arrow (1×; reads aligned with pbmm2 v1.0.0; SMRT Link) or filtered Illumina reads with POLCA (1×; part of MaSuRCA v4.0.5) (Zimin et al. 2013) followed by NextPolish (1×). The contig start positions in each polished assembly were adjusted using Circlator v1.5.5 (Hunt et al. 2015), followed by one final round of polishing using Arrow for the PacBio reads or NextPolish for the Illumina reads to ensure clean circularization. The quality and completeness of each genome assembly was evaluated using QUAST v5.0.2 (Gurevich et al. 2013) and BUSCO v3.0.1 (Enterobacteriales odb9 database) (Simão et al. 2015). The final genomes were annotated using the National Center for Biotechnology Information Prokaryotic Genome Annotation Pipeline (Tatusova et al. 2016) during submission of the closed genome sequences to GenBank, using plasmid names as described below.

The genome of each of the four strains consisted of a single circular chromosome and one circular extrachromosomal contig (i.e., plasmid) in *Et-C1* strains BHKY and BuffGH or five plasmids in *Et-melo* strains MDCuke and SCR3. To facilitate naming of these plasmids, homologous plasmids shared by multiple strains were identified by performing a BLASTn query of each contig against the complete assemblies of the other three strains; plasmids were considered homologous if their sequences shared >80% identity over >60% of the query length. Homologous plasmids were manually reoriented to start at the same arbitrarily selected intergenic region on the same strand, using SnapGene Viewer v5.3.2, and the reoriented plasmids were subsequently aligned using Mummer and progressiveMauve v20150226 (Darling et al. 2004) to confirm homology. Plasmids were named using the convention "pETR00X-y", with X reflecting a distinct plasmid and y denoting a strain-specific homolog. The four sequenced strains harbored a total of seven distinct plasmids, ranging in size from approximately 7 to 70 kb. A variant of the plasmid present in BHKY (pETR004-b) and BuffGH (pETR004-c) was present, albeit smaller, in MDCuke (pETR004-a) but was absent from SCR3 (Table 1). Of the remaining plasmids, three were shared by MDCuke and SCR3 (pETR001-a and b, pETR002-a and b, pETR005-a and b), one was unique to MDCuke (pETR003-a), and two were unique to SCR3 (pETR006-a, pETR007-a). The unique plasmid in MDCuke, pETR003-a, is homologous to the sequence that was reported in a draft MDCuke assembly as phage LS-2018a (59,759 bp) (Shapiro et al. 2018b), which was a partial concatamer (Thompson et al. 2019); pETR003-a represents the trimmed, circularized sequence of this putative temperate phage. Additional genome characteristics of these four strains are presented in Table 1.

Data Availability

The complete genome sequences and raw sequencing reads for BHKY, BuffGH, MDCuke, and SCR3 have been deposited in GenBank under BioProject accession number PRJNA788515. The GenBank accession numbers are listed in Table 1.

Author-Recommended Internet Resources

Bandage v0.8.1: https://rrwick.github.io/Bandage
BEDtools v2.27.1 tool 'getfasta': http://bedtools.readthedocs.io/en/latest
BLASTn: https://blast.ncbi.nlm.nih.gov/Blast.cgi
BUSCO v3.0.1: https://busco-archive.ezlab.org
Circlator v1.5.5: https://sanger-pathogens.github.io/circlator
FastQC v0.11.7: https://www.bioinformatics.babraham.ac.uk/projects/fastqc
Flye v2.9: https://github.com/fenderglass/Flye
MaSuRCA v4.0.5: https://github.com/alekseyzimin/masurca
Miniasm (v0.3): https://github.com/lh3/miniasm
Minipolish (v0.1.3): https://github.com/rrwick/Minipolish
Mummer v3.23: http://mummer.sourceforge.net

NextDenovo/NextPolish (v2.5.0/v1.4.0): https://github.com/Nextomics/NextDenovo
NanoFilt v2.8.0: https://github.com/wdecoster/nanofilt
NanoPlot v1.38.1: https://github.com/wdecoster/NanoPlot
NanoStat v1.5.0: https://github.com/wdecoster/nanostat
Porechop v0.2.4: https://github.com/rrwick/Porechop
progressiveMauve: http://darlinglab.org/mauve/mauve.html
QUAST v5.0: http://quast.sourceforge.net/quast
Raven v1.6.1: https://github.com/lbcb-sci/raven
Sambrook and Russell phenol-chloroform-based protocol: https://www.protocols.io/view/ultra-long-read-sequencing-protocol-for-rad004-mrxc57n
SnapGene Viewer v5.3.2: https://www.snapgene.com/snapgene-viewer
Trimmomatic: http://www.usadellab.org/cms/index.php?page=trimmomatic
Trycycler v0.4.1: https://github.com/rrwick/Trycycler
Unicycler v0.4.9: https://github.com/rrwick/Unicycler

Literature Cited

Bolger, A. M., Lohse, M., and Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Cavanagh, A., Hazzard, R., Adler, L. S., and Boucher, J. 2009. Using trap crops for control of *Acalymma vittatum* (Coleoptera: Chrysomelidae) reduces insecticide use in butternut squash. J. Econ. Entomol. 102:1101-1107.

Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394-1403.

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., and Van Broeckhoven, C. 2018. NanoPack: Visualizing and processing long-read sequencing data. Bioinformatics 34:2666-2669.

Delcher, A. L., Phillippy, A., Carlton, J., and Salzberg, S. L. 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 30:2478-2483.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29:1072-1075.

Hunt, M., Silva, N. D., Otto, T. D., Parkhill, J., Keane, J. A., and Harris, S. R. 2015. Circlator: Automated circularization of genome assemblies using long sequencing reads. Genome Biol. 16:294.

King, E. O., Ward, M. K., and Raney, D. E. 1954. Two simple media for the demonstration of pyocyanin and fluorescin. J. Lab. Clin. Med. 44:301-307.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. 2019. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37:540-546.

Lis, J. T., and Schleif, R. 1975. Size fractionation of double-stranded DNA by precipitation with polyethylene glycol. Nucleic Acids Res. 2:383-390.

Mayjonade, B., Gouzy, J., Donnadieu, C., Pouilly, N., Marande, W., Callot, C., Langlade, N., and Muños, S. 2016. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. Biotechniques 61:203-205.

Quinlan, A. R., and Hall, I. M. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

Saalau Rojas, E. S., Batzer, J. C., Beattie, G. A., Fleischer, S. J., Shapiro, L. R., Williams, M. A., Bessin, R., Bruton, B. D., Boucher, T. J., Jesse, L. C. H., and Gleason, M. L. 2015. Bacterial wilt of cucurbits: Resurrecting a classic pathosystem. Plant Dis. 99:564-574.

Saalau Rojas, E. S., Dixon, P. M., Batzer, J. C., and Gleason, M. L. 2013. Genetic and virulence variability among *Erwinia tracheiphila* strains recovered from different cucurbit hosts. Phytopathology 103:900-905.

Saalau Rojas, E. S., and Gleason, M. L. 2012. Epiphytic survival of *Erwinia tracheiphila* on muskmelon (*Cucumis melo* L.). Plant Dis. 96:62-66.

Sánchez, E. S., Hernández, E., Gleason, M. L., Batzer, J. C., Williams, M. A., Coolong, T., and Bessin, R. 2015. Optimizing rowcover deployment for managing bacterial wilt and using compost for organic muskmelon production. HortTechnology 25:762-768.

Shapiro, L. R., Andrade, A., Scully, E. D., Rocha, J., Paulson, J. N., and Kolter, R. 2018a. Draft genome sequence of an *Erwinia tracheiphila* isolate from an infected muskmelon (*Cucumis melo*). Microbiol. Resour. Announc. 7:e01058-18.

Shapiro, L. R., Paulson, J. N., Arnold, B. J., Scully, E. D., Zhaxybayeva, O., Pierce, N. E., Rocha, J., Klepac-Ceraj, V., Holton, K., and Kolter, R. 2018b. An introduced crop plant is driving diversification of the virulent bacterial pathogen *Erwinia tracheiphila*. MBio 9:e01307-18.

Shapiro, L. R., Scully, E. D., Roberts, D., Straub, T. J., Geib, S. M., Park, J., Stephenson, A. G., Salaau Rojas, E., Liu, Q., Beattie, G., Gleason, M., De Moraes, C. M., Mescher, M. C., Fleischer, S. G., Kolter, R., Pierce, N., and Zhaxybayeva, O. 2015. Draft genome sequence of *Erwinia tracheiphila*, an economically important bacterial pathogen of cucurbits. Genome Announc. 3:e00482-15.

Shapiro, L. R., Scully, E. D., Straub, T. J., Park, J., Stephenson, A. G., Beattie, G. A., Gleason, M. L., Kolter, R., Coelho, M. C., De Moraes, C. M., Mescher, M. C., and Zhaxybayeva, O. 2016. Horizontal gene acquisitions, mobile element proliferation, and genome decay in the host-restricted plant pathogen *Erwinia tracheiphila*. Genome Biol. Evol. 8:649-664.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-3212.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., and Ostell, J. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 44: 6614-6624.

Thompson, D. W., Casjens, S. R., Sharma, R., and Grose, J. H. 2019. Genomic comparison of 60 completely sequenced bacteriophages that infect *Erwinia* and/or *Pantoea* bacteria. Virology 535:59-73.

Vaser, R., and Šikić, M. 2021. Time- and memory-efficient genome assembly with Raven. Nat. Comput. Sci. 1:332-336.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27:737-746.

Vrisman, C. M., Deblais, L., Rajashekara, G., and Miller, S. A. 2016. Differential colonization dynamics of cucurbit hosts by *Erwinia tracheiphila*. Phytopathology 106:684-692.

Weber, D. C. 2018. Field attraction of striped cucumber beetles to a synthetic vittatalactone mixture. J. Econ. Entomol. 111:2988-2991.

Wick, R. R., and Holt, K. E. 2021. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. [Version 4; peer review: 4 approved] F1000 Res. 8:2138.

Wick, R. R., Judd, L. M., Cerdeira, L. T., Hawkey, J., Méric, G., Vezina, B., Wyres, K. L., and Holt, K. E. 2021. Trycycler: Consensus long-read assemblies for bacterial genomes. Genome Biol. 22:266.

Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Comput. Biol. 13:e1005595.

Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. 2015. Bandage: Interactive visualization of de novo genome assemblies. Bioinformatics 31:3350-3352.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. 2013. The MaSuRCA genome assembler. Bioinformatics 29:2669-2677.