AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Complete Genome Sequence of the *Arcobacter molluscorum* Type Strain LMG 25693

William G. Miller,[a] Emma Yee,[a] James L. Bono[b]

[a]U.S. Department of Agriculture, Produce Safety and Microbiology Research Unit, Agricultural Research Service, Albany, California, USA

[b]U.S. Department of Agriculture, Meat Safety and Quality Research Unit, Agricultural Research Service, Clay Center, Nebraska, USA

ABSTRACT As components of freshwater and marine microflora, *Arcobacter* spp. are often recovered from shellfish, such as mussels, clams, and oysters. *Arcobacter molluscorum* was isolated from mussels from the Ebro Delta in Catalonia, Spain. This article describes the whole-genome sequence of the *A. molluscorum* strain LMG 25693^T (= F98-3^T = CECT 7696^T).

Members of the genus *Arcobacter* are often recovered from shellfish (1–7). The prevalence of *Arcobacter* species in environmental waters (8) suggests that contamination of shellfish by these organisms might be the result of filter feeding-associated bioaccumulation, with this contamination potentially resulting in human illness following the consumption of raw or partially cooked shellfish. *Arcobacter molluscorum* was isolated from farmed shellfish harvested in Catalonia, Spain (4). In this article, we report the first closed genome sequence of the *A. molluscorum* type strain LMG 25693 (= F98-3^T = CECT 7696^T), isolated in 2009 from farmed mussels from the Ebro Delta in Catalonia, Spain.

The genome of *A. molluscorum* strain LMG 25693^T was completed using the Roche GS FLX+, Illumina HiSeq, and PacBio RS II next-generation sequencing platforms. Genomic DNA was isolated with the Wizard genomic DNA purification kit (Promega, Madison, WI) using a loop (~5 μl) of cells taken from cultures grown (aerobic environment, 48 h, 30°C) on anaerobe basal agar (Oxoid) amended with 5% horse blood. Shotgun and paired-end Roche 454 libraries were constructed following the manufacturer's protocols, and 454 sequencing was performed using the Titanium chemistry and standard methods. PacBio SMRTbell libraries were prepared from 10 μg of genomic DNA using the standard 20-kb PacBio protocol (9). Single-molecule real-time (SMRT) cell sequencing was performed using standard protocols, the 20-kb libraries, P6-C4 sequencing chemistry, and the 360-min data collection mode. Illumina HiSeq reads were obtained from SeqWright (Houston, TX). Shotgun and paired-end Roche 454 reads were assembled using Newbler v. 2.6 (Roche) and default parameters into 88 total contigs; 5 low-quality contigs consisting of <100 reads were deleted. PacBio reads were assembled with RS Hierarchical Genome Assembly Process (HGAP) v. 3 (Pacific Biosciences) with default settings, which yielded a single chromosomal contig that was polished, using the RS.Resequencing.1 module (Pacific Biosciences) with default parameters, and circularized. Reads were quality controlled within the Newbler or RS HGAP assemblers; 99.8% to 99.99% of the bases in the assembled 454 and Illumina contigs had base call quality scores of 40 (Table 1). The custom Perl script contig_extender3 (10) was used to order and orient the 454 contigs into a single circular sequence. Verification of this 454 contig order was performed through a BLASTN analysis of these contigs using the PacBio contig as a reference. The 55 unique 454 contigs and the PacBio contig were assembled together using SeqMan Pro v. 8.0

Received 18 September 2018 **Accepted** 26 September 2018 **Published** 25 October 2018

Citation Miller WG, Yee E, Bono JL. 2018. Complete genome sequence of the *Arcobacter molluscorum* type strain LMG 25693. Microbiol Resour Announc 7:e01293-18. https://doi.org/10.1128/MRA.01293-18.

Editor John J. Dennehy, Queens College

Address correspondence to William G. Miller, william.miller@ars.usda.gov.

TABLE 1 Sequencing metrics and genomic data for *A. molluscorum* strain LMG 25693^T

Feature	Value(s)[a]
Sequencing metrics	
454 (shotgun) platform	
No. of reads	177,873
No. of bases	73,714,660
Average length (bases)	414.4
Coverage (×)	26.3
454 (paired-end) platform	
No. of reads	150,593
No. of bases	46,384,064
Average length (bases)	308.0
Coverage (×)	16.6
Illumina HiSeq 2000 platform	
No. of reads	25,306,576
No. of bases	2,530,657,600
Average length (bases)	100
Coverage (×)	903.6
PacBio platform	
No. of reads	129,047
No. of bases	399,548,656
Average length (bases)	3,096.1[b]
Coverage (×)	142.7
Newbler metrics[c]	
N50ContigSize (454) (bases)	90,324
Q40PlusBases (454) (%)	99.84
N50ContigSize (HiSeq pool 1) (bases)	78,972
Q40PlusBases (HiSeq pool 1) (%)	99.99
N50ContigSize (HiSeq pool 2) (bases)	90,503
Q40PlusBases (HiSeq pool 2) (%)	99.96
N50ContigSize (HiSeq pool 3) (bases)	79,027
Q40PlusBases (HiSeq pool 3) (%)	99.97
Genomic data	
Chromosome	
Size (bp)	2,800,582
G+C content (%)	26.25
No. of CDS[d]	2,666
Assigned function (% CDS)	1,044 (39.2)
General function annotation (% CDS)	995 (37.3)
Domain/family annotation only (% CDS)	199 (7.5)
Hypothetical (% CDS)	428 (16.1)
Pseudogenes	31
Genomic islands/CRISPR	
No. of genetic islands	3
No. of CDS in genetic islands	71, [1]
CRISPR-Cas loci	I-B, [III-A]
Gene content/pathways	
IS elements, mobile elements, or tranposases	3 (IS*1595*); 1, [1] (other)
Signal transduction	
Che proteins	*cheABDRVW*(*Y*)$_2$
No. of methyl-accepting chemotaxis proteins	26
No. of response regulators	57
No. of histidine kinases	62
No. of response regulator/histidine kinase fusions	7
No. of diguanylate cyclases	17
No. of diguanylate phosphodiesterases (HD-GYP, EAL)	4, 5
No. of diguanylate cyclase/phosphodiesterases	8
No. of other	11
Motility	
Flagellin genes	*fla1* to *fla6*
Restriction/modification	
No. of type I systems (*hsd*)	1
No. of type II systems	1, [1]
No. of type III systems	0
Transcription/translation	
No. of transcriptional regulatory proteins	64
Non-ECF[e] σ factors	σ^{54}, σ^{70}

(Continued on next page)

TABLE 1 (Continued)

Feature	Value(s)[a]
No. of ECF σ factors	0
No. of tRNAs	56
No. of ribosomal loci[f]	3 (A), 3 (B)
CO dehydrogenase (*coxSLF*)	Yes
Ethanolamine utilization (*eutBCH*)	Yes
Nitrogen fixation (*nif*)	Yes
Osmoprotection	$BCCT_3$, *ectABC*
Pyruvate → acetyl-CoA	
Pyruvate dehydrogenase (E1/E2/E3)	Yes
Pyruvate:ferredoxin oxidoreductase	*por*
Urease	*ureAB*
Vitamin B_{12} biosynthesis	Yes

[a]Numbers in square brackets indicate pseudogenes or fragments.
[b]Maximum length, 25,747 bases.
[c]Features and values taken from largeContigMetrics within 454NewblerMetrics.txt for each assembly. Large contigs were defined as ≥500 bases. Due to the large number of HiSeq reads, the total reads were split into three pools and assembled independently.
[d]Numbers do not include pseudogenes; CDS, coding sequences.
[e]ECF, extracytoplasmic function.
[f]A: 16S-$tRNA_{Ile}$-$tRNA_{Ala}$-23S-5S; B: 16S-23S-5S.

(DNASTAR, Madison, WI), with the remaining 28 contigs that represent repeat regions added to the assembly manually at two or more locations. This assembly was confirmed using an optical restriction map (restriction enzyme XbaI; OpGen, Gaithersburg, MD). Verification and error correction of base calls within the composite 454/PacBio assembly were performed using the HiSeq reads. These reads were assembled *de novo* within Newbler using the same parameters as with the 454 reads; small contigs represented by <20 reads were deleted. The remaining contigs were assembled into the SeqMan 454/PacBio assembly described above, with base calls adjusted to the Illumina consensus sequence. Single nucleotide polymorphisms within the repeat contigs and sequences between the Illumina contigs were assessed/verified by assembling the Illumina reads onto these regions within Geneious v. 8.1 (Biomatters, Auckland, NZ) and using the "find variations/SNPs" module, with a default minimum variant frequency parameter of 0.3. The final coverage across the genome was 1,089×.

A. molluscorum strain LMG 25693^T has a circular genome of 2,800,582 bp with an average G+C content of 26.25%. Protein-, rRNA-, and tRNA-encoding genes were identified and annotated as described (11, 12). Briefly, putative coding sequences (CDSs), tRNA/transfer-messenger RNA (tmRNA) genes, and rRNA loci were identified using GeneMark, ARAGORN, and RNAmmer, respectively (13–15). The genome sequence and the CDS coordinates from GeneMark were used to create a preliminary GenBank-formatted file which was entered into Artemis v. 16 (16) to identify putative pseudogenes and genes missed in the original GeneMark analysis and to manually curate the start codon of each putative CDS. Initial annotation was accomplished by comparing the proteome of strain LMG 25693^T to proteomes derived from other *Arcobacter* genomes (primarily *A. butzleri* strain RM4018 and *A. nitrofigilis* [GenBank accession numbers CP000361 and CP001999, respectively]) and to proteins in the NCBI nonredundant (nr) database using BLASTP. Annotation was further refined, e.g., through an analysis of Pfam motifs (17) and a BLASTP analysis that utilized a larger custom protein database that also included proteomes from all current completed *Campylobacter* genomes.

The LMG 25693^T genome is predicted to encode 2,666 putative protein-coding genes and 31 pseudogenes. Additionally, the LMG 25693^T genome contains 56 tRNA-encoding genes and 6 rRNA operons; however, 3 of these rRNA operons do not contain the isoleucyl-tRNA or alanyl-tRNA genes that are commonly found in other rRNA loci. Three genomic islands were identified in the LMG 25693^T genome; one genomic island is a putative integrated plasmid containing genes for a P-type type IV conjugative transfer system, while a second 28-kb island putatively encodes a type VI secretion

system. The LMG 25693^T genome also contains a type I-B CRISPR-Cas system. A second CRISPR-Cas system (type III-A) was identified; however, although this locus contains the *cas6*, *csm2*, *csm3*, *csm4*, and *csm5* genes, it does not contain *cas1* or *cas2*, and the *cas10* gene is presumably nonfunctional. No plasmids were identified in the strain LMG 25693^T genome.

Data availability. The complete genome sequence of *A. molluscorum* strain LMG 25693^T has been deposited in GenBank under the accession number CP032098. HiSeq, 454, and PacBio sequencing reads have been deposited in the NCBI Sequence Read Archive (SRA; accession number SRP155187).

ACKNOWLEDGMENTS

This work was funded by the United States Department of Agriculture, Agricultural Research Service, Current Research Information System (CRIS) projects 2030-42000-230-047, 2030-42000-230-051, and 3040-42000-015-00D.

We thank Maria Figueras for providing *A. molluscorum* strain LMG 25693^T.

REFERENCES

1. Collado L, Cleenwerck I, Van Trappen S, De Vos P, Figueras MJ. 2009. *Arcobacter mytili* sp. nov., an indoxyl acetate-hydrolysis-negative bacterium isolated from mussels. Int J Syst Evol Microbiol 59:1391–1396. https://doi.org/10.1099/ijs.0.003749-0.
2. Collado L, Guarro J, Figueras MJ. 2009. Prevalence of *Arcobacter* in meat and shellfish. J Food Prot 72:1102–1106. https://doi.org/10.4315/0362-028X-72.5.1102.
3. Dieguez AL, Balboa S, Magnesen T, Romalde JL. 2017. *Arcobacter lekithochrous* sp. nov., isolated from a molluscan hatchery. Int J Syst Evol Microbiol 67:1327–1332. https://doi.org/10.1099/ijsem.0.001809.
4. Figueras MJ, Collado L, Levican A, Perez J, Solsona MJ, Yustes C. 2011. *Arcobacter molluscorum* sp. nov., a new species isolated from shellfish. Syst Appl Microbiol 34:105–109. https://doi.org/10.1016/j.syapm.2010.10.001.
5. Figueras MJ, Levican A, Collado L, Inza MI, Yustes C. 2011. *Arcobacter ellisii* sp. nov., isolated from mussels. Syst Appl Microbiol 34:414–418. https://doi.org/10.1016/j.syapm.2011.04.004.
6. Levican A, Collado L, Aguilar C, Yustes C, Dieguez AL, Romalde JL, Figueras MJ. 2012. *Arcobacter bivalviorum* sp. nov. and *Arcobacter venerupis* sp. nov., new species isolated from shellfish. Syst Appl Microbiol 35:133–138. https://doi.org/10.1016/j.syapm.2012.01.002.
7. Levican A, Collado L, Yustes C, Aguilar C, Figueras MJ. 2014. Higher water temperature and incubation under aerobic and microaerobic conditions increase the recovery and diversity of *Arcobacter* spp. from shellfish. Appl Environ Microbiol 80:385–391. https://doi.org/10.1128/AEM.03014-13.
8. Ramees TP, Dhama K, Karthik K, Rathore RS, Kumar A, Saminathan M, Tiwari R, Malik YS, Singh RK. 2017. *Arcobacter*: an emerging food-borne zoonotic pathogen, its public health concerns and advances in diagnosis and control—a comprehensive review. Vet Q 37:136–161. https://doi.org/10.1080/01652176.2017.1323355.
9. PacBio. 2015. Procedure and checklist: 20 kb template preparation using BluePippin size-selection system. https://www.pacb.com/wp-content/uploads/2015/09/Procedure-Checklist-20-kb-Template-Preparation-Using-BluePippin-Size-Selection.pdf. Accessed 24 September 2018.
10. Miller WG, Yee E, Bono JL. Complete genome sequence of the *Arcobacter halophilus* type strain CCUG 53805. Microbiol Resour Announc, in press.
11. Miller WG, Yee E, Chapman MH, Smith TP, Bono JL, Huynh S, Parker CT, Vandamme P, Luong K, Korlach J. 2014. Comparative genomics of the *Campylobacter lari* group. Genome Biol Evol 6:3252–3266. https://doi.org/10.1093/gbe/evu249.
12. Miller WG, Yee E, Bono JL. 2018. Complete genome sequence of the Arcobacter bivalviorum type strain LMG 26154. Microbiol Resour Announc 7: e01076-18. https://mra.asm.org/content/7/12/e01076-18/article-info.
13. Besemer J, Borodovsky M. 2005. GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 33:W451–W454. https://doi.org/10.1093/nar/gki487.
14. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108. https://doi.org/10.1093/nar/gkm160.
15. 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.
16. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944–945. https://doi.org/10.1093/bioinformatics/16.10.944.
17. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD. 2012. The Pfam protein families database. Nucleic Acids Res 40:D290–D301. https://doi.org/10.1093/nar/gkr1065.