AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Genome Sequence of a Microvirus Recovered from Wastewater in Arizona, USA, in October 2020, Encodes a Previously Undescribed DNA-Binding Protein

Abriana Smith,[a] Nicole Kaiser,[a] Allan Yanez,[b] Tyler Perleberg,[b] Amir Elyaderani,[a] Peter Skidmore,[a] Sangeet Adhikari,[b,c] Erin M. Driver,[b] Rolf U. Halden,[b,c,d] Arvind Varsani,[e] Matthew Scotch,[a,b] Temitope O. C. Faleye[b]

[a]College of Health Solutions, Arizona State University, Tempe, Arizona, USA
[b]Biodesign Center for Environmental Health Engineering, Biodesign Institute, Arizona State University, Tempe, Arizona, USA
[c]School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, Arizona, USA
[d]OneWaterOneHealth, Arizona State University Foundation, Tempe, Arizona, USA
[e]Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, Arizona, USA

Abriana Smith and Nicole Kaiser contributed equally to this work. The order of names was determined alphabetically in the order of their first names.

ABSTRACT We describe the genome of Microvirus-AZ-2020, which was identified from wastewater in Arizona, USA, in October 2020. Microvirus-AZ-2020 belongs to subfamily *Gokushovirinae* and contains six (five known and one hypothetical) open reading frames (ORFs), each with >40 codons. HHPred analysis and Colabfold structure prediction suggest that the hypothetical ORF encodes a previously undescribed putative DNA-binding protein.

Microviruses are bacteriophages in the *Microviridae* family and are circular single-stranded DNA viruses with icosahedral capsids. Metagenomic studies have identified microviruses in a range of environments (1, 2). We describe the genome of Microvirus-AZ-2020, which was identified from wastewater collected in Arizona, USA, in October 2020.

The 2-L postfiltration sample (450 nM) was concentrated to 2 mL by ultrafiltration using a 10,000-molecular-weight-cutoff centrifugal filter. The concentrate was subjected to nucleic acid extraction (QIAamp minikit), pan-enterovirus reverse transcription (RT)-PCR (3), library preparation (KAPA HyperPlus library kit), and paired-end (2 × 250-bp) sequencing (MiSeq system; Illumina). Raw reads were trimmed and *de novo* assembled using Trimmomatic v0.36 and metaSPAdes v3.15.3, respectively, on the KBase platform (4). Contigs were identified using a BLASTn search of the GenBank database (5). The proportion of reads mapped to Microvirus-AZ-2020 (template) and the depth of coverage were determined using Bowtie2 v2.3.2 (3). The Microvirus-AZ-2020 open reading frames (ORFs) were predicted using Prokka v1.0.0 (6) and DNA Master v5.23.6 (7). Functional annotation of the predicted ORFs was performed via a BLASTp search of the nonredundant protein sequence database using the BLAST-All-Genes option in DNA Master. Unannotated ORFs were subjected to HHPred analysis (8, 9) and protein structure prediction using ColabFold (10), which combines the fast homology search of MMseqs2 with AlphaFold2. DNA-binding residues were predicted using the DRNApred server (11). All software was used with default parameters unless otherwise specified. Primers MazF (5′-GTGGCGAAGCCGGCATGGGTTGTTAGGG-3′) and MazR (5′-GTGGCGAAGCCGGCATGGGTTGTTAGGGAGAAACCC-3′) (Fig. 1A) were used to confirm the presence of virus in the sample by PCR using Phusion green master mix with the following reaction conditions: 94°C for 3 min, 40 cycles of 94°C for 30 s, 55°C for 30 s, and 68°C for 6 min, and finally 68°C for 10 min.

Microvirus-AZ-2020 (4,665 nucleotides [nt] [GC content, 51%]) was *de novo* assembled from 2,561 reads (0.07% of the 3,913,700 trimmed reads [depth of coverage, 111×]), circularized via terminal redundancy, and determined by BLASTn to be most similar to *Apis*

Editor John J. Dennehy, Queens College CUNY

Address correspondence to Temitope O. C. Faleye, tfaleye@asu.edu.
The authors declare a conflict of interest.

Received 29 May 2022
Accepted 18 August 2022
Published 31 August 2022

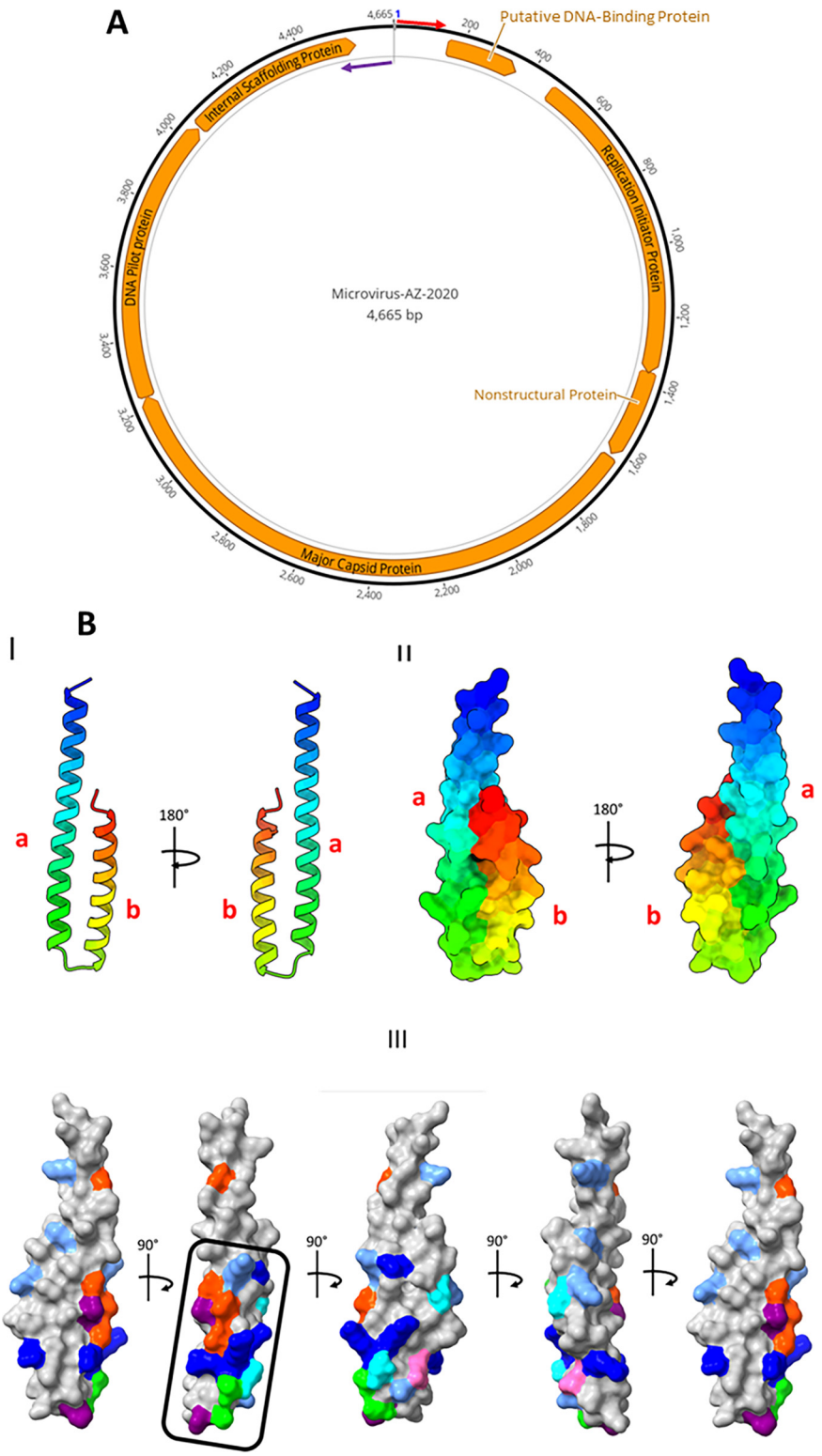

FIG 1 (A) Genome map of Microvirus-AZ-2020. The binding site of the primers used to amplify the complete genome is indicated by colored arrows. (B) Colabfold-predicted structure of the putative protein in panel A. The structure was viewed and annotated in ChimeraX. I is shown in ribbon view, while II and III are shown in surface view. I and II show the amino terminus to the carboxyl terminus from blue to red. In I and II, a and b refer to the two alpha-helices from amino to carboxyl terminal, respectively. III shows the result of DNA-binding residue prediction layered on the predicted three-dimensional structure of the molecule (blue, cornflower blue, purple, pink, cyan, green, and orange indicate amino acid residues R, K, H, Q, N, S, and T, respectively). The region highlighted with a black box in III shows clustering of some predicted DNA-binding residues.

TABLE 1 Details of predicted ORFs in Microvirus-AZ-2020

ORF no.	Nucleotide positions of ORF	Length (nt)	Length (amino acids)	Start codon	Stop codon	GenBank accession no. of most similar protein sequence (species)	Alignment (%)	Identity (%)	BLASTp-predicted function
1	152–358	207	68	ATG	TGA	AZL82829 (*Apis mellifera*-associated microvirus 13)	100	31.34	Hypothetical protein
2	468–1361	894	297	GTG	TGA	AXH74636 (*Microviridae* sp.)	100	61.74	Replication initiator protein
3	1358–1612	255	84	ATG	TGA	AZL82825 (*Apis mellifera*-associated microvirus 13)	100	54.32	Nonstructural protein
4	1621–3234	1614	537	ATG	TGA	AZL82828 (*Apis mellifera*-associated microvirus 13)	100	70.30	Major capsid protein
5	3237–4046	810	269	ATG	TGA	AZL82826 (*Apis mellifera*-associated microvirus 13)	92.2	50.96	DNA pilot protein
6	4051–4557	507	168	ATG	TAA	AZL82827 (*Apis mellifera*-associated microvirus 13)	94.9	57.05	Internal scaffolding protein

mellifera-associated microvirus (subfamily *Gokushovirinae*) (GenBank accession MH992184 [12]) (query coverage, 77%; pairwise identity, 67.75%; E score, 0.0). Both Prokka (Fig. 1A) and DNA Master (Table 1) predicted six ORFs with >40 codons. A BLASTp search of the nonredundant protein sequence database annotated five of the predicted ORFs as encoding replication initiator protein, nonstructural protein, major capsid protein, DNA pilot protein, and internal scaffolding protein (Table 1 and Fig. 1A). No function could be assigned to the sixth (hypothetical) ORF using BLASTp. However, 8 of its top 10 HHPred analysis hits were transcription regulation proteins (probability, 67.2% to 83.35%). ColabFold structure prediction showed that the protein has a helix-turn-helix motif (Fig. 1BI), suggesting that it is a potential DNA-binding protein. When it was layered on the three-dimensional structure, DRNApred predicted that the DNA-binding residues spanned both α-helices but clustered toward the lower half (from the amino end) of helix a (Fig. 1BIII).

We describe the genome of Microvirus-AZ-2020, a microvirus recovered from wastewater in Arizona, USA, in October 2020 that encodes a previously undescribed putative DNA-binding protein. Surveillance of microviruses is needed to improve our understanding of their diversity and unexplored protein repertoire.

Data availability. The mapped reads and microvirus genome described in this study have been deposited in the SRA and GenBank under accession numbers SRR18497324 and ON111452, respectively.

ACKNOWLEDGMENTS

We thank the City of Tempe wastewater management authorities for help with sample collection from the city sewers. We also thank the Genomics Core at the Biodesign Institute, Arizona State University, for help with library preparation and Illumina sequencing.

The research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award U01LM013129 to R.U.H., M.S., and A.V.

The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

E.M.D. and R.U.H. are cofounders of AquaVitas, LLC (Scottsdale, AZ), an Arizona State University start-up company providing commercial services in wastewater-based epidemiology. R.U.H. is the founder of OneWaterOneHealth, a nonprofit project of the Arizona State University Foundation.

REFERENCES

1. Roux S, Krupovic M, Poulet A, Debroas D, Enault F. 2012. Evolution and diversity of the *Microviridae* viral family through a collection of 81 new complete genomes assembled from virome reads. PLoS One 7:e40418. https://doi.org/10.1371/journal.pone.0040418.
2. Kirchberger PC, Martinez ZA, Ochman H. 2022. Organizing the global diversity of microviruses. mBio 13:e0058822. https://doi.org/10.1128/mbio.00588-22.
3. Faleye TOC, Driver E, Bowes D, Adhikari S, Adams D, Varsani A, Halden RU, Scotch M. 2021. Pan-enterovirus amplicon-based high-throughput sequencing detects the complete capsid of a EVA71 genotype C1 variant via wastewater-based epidemiology in Arizona. Viruses 13:74. https://doi.org/10.3390/v13010074.
4. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez F, Canon S, Sneddon MW, Henderson ML, Riehl WJ, Murphy-Olson D, Chan SY, Kamimura RT, Kumari S, Drake MM, Brettin TS, Glass EM, Chivian D, Gunter D, Weston DJ, Allen BH, Baumohl J, Best AA, Bowen B, Brenner SE, Bun CC, Chandonia JM, Chia JM, Colasanti R, Conrad N, Davis JJ,

Davison BH, DeJongh M, Devoid S, Dietrich E, Dubchak I, Edirisinghe JN, Fang G, Faria JP, Frybarger PM, Gerlach W, Gerstein M, Greiner A, Gurtowski J, Haun HL, He F, Jain R, et al. 2018. KBase: the United States Department of Energy Systems Biology Knowledgebase. Nat Biotechnol 36:566–569. https://doi.org/10.1038/nbt.4163.
5. 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.
6. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
7. Pope WH, Jacobs-Sera D. 2018. Annotation of bacteriophage genome sequences using DNA Master: an overview. Methods Mol Biol 1681:217–229. https://doi.org/10.1007/978-1-4939-7343-9_16.
8. Söding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960. https://doi.org/10.1093/bioinformatics/bti125.
9. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A completely reimplemented MPI Bioinformatics toolkit with a new HHpred server at its core. J Mol Biol 430:2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007.
10. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. Nat Methods 19:679–682. https://doi.org/10.1038/s41592-022-01488-1.
11. Yan J, Kurgan L. 2017. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res 45:e84. https://doi.org/10.1093/nar/gkx059.
12. Kraberger S, Cook CN, Schmidlin K, Fontenele RS, Bautista J, Smith B, Varsani A. 2019. Diverse single-stranded DNA viruses associated with honey bees (*Apis mellifera*). Infect Genet Evol 71:179–188. https://doi.org/10.1016/j.meegid.2019.03.024.