AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

Genomics and Proteomics | Announcement

Draft genome of *Microsporidia* sp. MB—a malaria-blocking microsporidian symbiont of the *Anopheles arabiensis*

Lilian Mbaisi Ang'ang'o,[1] Jacqueline Wahura Waweru,[2] Edward Edmond Makhulu,[2] Anne Wairimu,[2] Fidel Gabriel Otieno,[2] Thomas Onchuru,[2] Özlem Tastan Bishop,[1] Jeremy Keith Herren[2]

AUTHOR AFFILIATIONS See affiliation list on p. 3.

ABSTRACT We report the draft whole-genome assembly of *Microsporidia* sp. MB, a symbiotic malaria-transmission-blocking microsporidian isolated from *Anopheles arabiensis* in Kenya. The whole-genome sequence of *Microsporidia* sp. MB has a length of 5,908,979 bp, 2,335 contigs, and an average GC content of 31.12%.

KEYWORDS microsporidia, anopheles, genomes, endosymbionts

Microsporidia are microscopic, obligate intracellular eukaryotes that widely infect both vertebrates and invertebrates (1–7). *Microsporidia* sp. MB is a species of microsporidia that infects *Anopheles* mosquitoes and has been identified as a potential malaria transmission-blocking agent, as it can significantly reduce the vectorial capacity of *Anopheles* (8). Moreover, it exhibits positive effects on the fitness of its host, contributing to its spread in host populations (8, 9). Its unique characteristics and life cycle adaptations make it an intriguing subject for research in mosquito-borne disease control (10). We aimed to sequence and assemble the genome of this important symbiont isolated from *Anopheles arabiensis* mosquitoes in Kenya.

Gravid female mosquitoes were collected in Ahero (34.9190°W, −0.1661°N), Western Kenya and used to set up isofemale family lines. Genomic DNA was extracted from dissected ovaries of *Microsporidia* sp. MB-infected F1 progenies using the protein precipitation extraction protocol, as previously described, and screened for the symbiont using MB18S primers quantitative PCR assays (8). Highly infected samples were selected for sequencing after assessing DNA quality using Qubit Fluorometric Quantitation (ThermoFisher Scientific, Waltham, USA). Paired short-insert libraries were prepared using KAPA HiFi HotStart Library Amp Kit and sequenced with DNBSeq technology (2 × 150 bp reads) at BGI Genomics (https://www.bgi.com/global), generating a total of 326,181,200 raw paired-end reads. SOAPnuke v2.1.8 was employed to filter the reads using filtering parameters: "*-n 0.001 L 10 -q 0.5 --adaMR 0.25 --polyX 50 –minReadLen 100*" (11). FastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (12) and MultiQC v1.12 (13) were used for quality assessment. The host reads were removed by mapping to the reference genomes of *A. arabiensis* (GenBank: GCA_000349185.1) and *A. gambiae s.s.* (GenBank: GCA_000005575.1) from the NCBI RefSeq database (release 219) (http://www.ncbi.nlm.nih.gov/RefSeq/) (14, 15) using the Burrows-Wheeler Aligner (BWA) v0.7.17 (https://github.com/lh3/bwa) (16). Samtools v1.3.1 (https://github.com/samtools/) (17) was used to filter out host-mapped reads. Kraken2 v2.0.8 (18) was applied to remove bacterial contaminants using the minikraken_8 GB_20200312 database. The clean reads were *de novo* assembled using Unicycler v0.4.9 (19), and a megablast search was conducted against microsporidia proteins to remove non-target contigs. The raw reads were reassembled to the cleaned assembly using BWA-MEM v0.7.17 (16) generating a consensus assembly. A remote BLAST against the NCBI nt database was used to identify contigs with high similarity to microsporidia. Gene

Editor Vincent Michael Bruno, University of Maryland School of Medicine, Baltimore, Maryland, USA

Address correspondence to Jeremy Keith Herren, jherren@icipe.org.

The authors declare no conflict of interest.

See the funding table on p. 3.

Received 29 September 2023
Accepted 9 March 2024
Published 21 March 2024

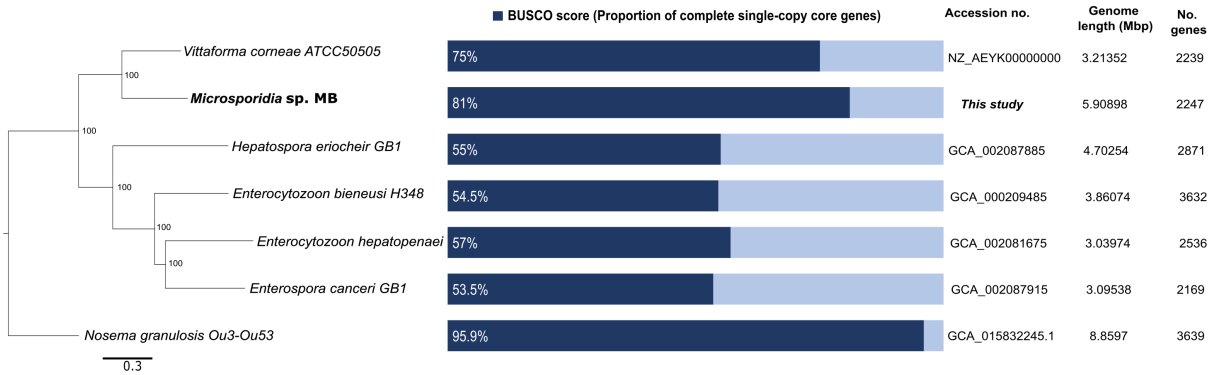

FIG 1 Phylogenomic analysis of *Microsporidia* sp. MB alongside genome assembly statistics comparison among the Enterocytozoonidae group of which five genomes have been fully sequenced. *Nosema granulosis* from the Nosematidae group of Microsporidia was used as an outgroup. The tree was constructed using maximum likelihood with FastTree v2 (26) based on the 399 single-copy orthologous genes found in all species using OrthoFinder v2.5.4 (25). Protein sequences were aligned with MAFFT v7 (27) using default options. Identified genes and species trees were generated on OrthoFinder v2.5.4 (28) and visualized on Dendroscope v3.8.4 (29). The phylogenomic tree reveals the close relationship between *Microsporidia* sp. MB and *Vittaforma corneae*.

TABLE 1 *Microsporidia* sp. MB genome assembly statistics

Metric	
Assembly size (bp)	5,908,979
Number of contigs	2,335
N_{50} (bp)	5,000
GC content (%)	31.12
The proportion of repeats (%)	0.57
Number of predicted genes	2,247
Gene density (genes/kb)	0.363
Mean CDS length (bp)	1,108
BUSCO (n = 600)	
No. (%) of complete genes	486 (81.0)
No. (%) of complete and single-copy genes	485 (80.8)
Number (%) of complete and duplicated genes	1 (0.2)
Number (%) of fragmented genes	12 (2.0)
Number (%) of missing genes	102 (17.0)
GenBank accession number	JAVKTW000000000
SRA accession number	SRR25938329

prediction was performed using GeneMarkS v4.3 (20) (intronless eukaryotic mode), and RepeatModeler v2.0.4 (http://www.repeatmasker.org) (21) used to identify repeats in the assembly (RepBaseRepeatMaskerEdition-20181026). Genome completeness was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.4.3 (22) against the microsporidia_odb10 database (n = 600) (23) indicating 81% completeness. Quality assessment and genome statistics were determined using QUAST v5.2.0 (24), revealing a total genome size of 5.90898 Mb spanning 2,335 contigs, with an N_{50} of 5,000 bp (Table 1).

Phylogenomic analysis using OrthoFinder v2.5.4 (25) showed *Microsporidia* sp. MB is closely related to *Vittaforma corneae* within the Enterocytozoonidae group, consistent with previous taxonomic classification based on SSU rRNA (4, 8, 10), and contained the highest proportion of core BUSCO genes (Fig. 1). Default parameters were used for all software except where otherwise noted.

ACKNOWLEDGMENTS

This work was carried out with the financial support of the Organization for Women in Science for the Developing World (OWSD); the Swedish International Development

Cooperation Agency (SIDA); Open Philanthropy (SYMBIOVECTOR Track A); the Bill and Melinda Gates Foundation (INV0225840); The Children's Investment Fund Foundation (SMBV-FFT), the Swiss Agency for Development and Cooperation (SDC); the Australian Centre for International Agricultural Research (ACIAR); the Federal Democratic Republic of Ethiopia; and the Government of the Republic of Kenya. The views expressed herein do not necessarily reflect the official opinion of the donors.

AUTHOR AFFILIATIONS

[1]Research Unit in Bioinformatics (RUBi), Department of Biochemistry and Microbiology, Rhodes University, Makhanda, Eastern Cape, South Africa
[2]International Centre of Insect Physiology and Ecology (icipe), Nairobi, Kenya

AUTHOR ORCIDs

Lilian Mbaisi Ang'ang'o http://orcid.org/0000-0001-8573-5417
Jeremy Keith Herren http://orcid.org/0000-0003-2239-7275

FUNDING

Funder	Grant(s)	Author(s)
Bill and Melinda Gates Foundation (GF)	INV0225840	Jeremy Keith Herren
Organization for Women in Science for the Developing World (OWSD)		Lilian Mbaisi Ang'ang'o
Open Philanthropy Project	SYMBIOVECTOR TRACK A	Jeremy Keith Herren
Children's Investment Fund Foundation (CIFF)	SMBV-FFT	Jeremy Keith Herren
Swedish International Development Cooperation Agency		Jeremy Keith Herren
The Swiss Agency for Development and Cooperation		Jeremy Keith Herren
The Australian Centre for International Agricultural Research		Jeremy Keith Herren

AUTHOR CONTRIBUTIONS

Lilian Mbaisi Ang'ang'o, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft | Jacqueline Wahura Waweru, Data curation, Methodology, Validation, Writing – review and editing | Edward Edmond Makhulu, Data curation, Methodology, Writing – review and editing | Anne Wairimu, Methodology | Fidel Gabriel Otieno, Methodology | Thomas Onchuru, Conceptualization, Investigation, Methodology, Project administration, Supervision | Özlem Tastan Bishop, Conceptualization, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – review and editing | Jeremy Keith Herren, Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review and editing

DATA AVAILABILITY

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession number JAVKTW000000000. The version described in this paper is the first version, JAVKTW010000000. The raw reads have been deposited at SRA under the accession SRR25938329.

REFERENCES

1. Wadi L, Reinke AW. 2020. Evolution of microsporidia: an extremely successful group of eukaryotic intracellular parasites. PLoS Pathog 16:e1008276. https://doi.org/10.1371/journal.ppat.1008276
2. Weiss LM, Reinke AW. 2022. Microsporidia: current advances in biology. Springer Nature Switzerland, Cham, Switzerland.
3. Stentiford GD, Feist SW, Stone DM, Bateman KS, Dunn AM. 2013. Microsporidia: diverse, dynamic, and emergent pathogens in aquatic systems. Trends Parasitol 29:567–578. https://doi.org/10.1016/j.pt.2013.08.005
4. Park E, Poulin R. 2021. Revisiting the phylogeny of microsporidia. Int J Parasitol 51:855–864. https://doi.org/10.1016/j.ijpara.2021.02.005
5. Didier ES. 2005. Microsporidiosis: an emerging and opportunistic infection in humans and animals. Acta Trop 94:61–76. https://doi.org/10.1016/j.actatropica.2005.01.010
6. Agnew P, Becnel JJ, Ebert D, Michalakis Y. 2003. Symbiosis of Microsporidia and insects, p 145–163. In Insect Symbiosis. CRC Press.
7. Becnel JJ, Andreadis TG. 2014. Microsporidia in insects, p 521–570. In Microsporidia: Pathogens of opportunity. Wiley Online Library.
8. Herren JK, Mbaisi L, Mararo E, Makhulu EE, Mobegi VA, Butungi H, Mancini MV, Oundo JW, Teal ET, Pinaud S, Lawniczak MKN, Jabara J, Nattoh G, Sinkins SP. 2020. A microsporidian impairs *Plasmodium falciparum* transmission in *Anopheles arabiensis* mosquitoes. Nat Commun 11:2187. https://doi.org/10.1038/s41467-020-16121-y
9. Nattoh G, Maina T, Makhulu EE, Mbaisi L, Mararo E, Otieno FG, Bukhari T, Onchuru TO, Teal E, Paredes J, Bargul JL, Mburu DM, Onyango EA, Magoma G, Sinkins SP, Herren JK. 2021. Horizontal transmission of the symbiont *Microsporidia* MB in *Anopheles arabiensis*. Front Microbiol 12. https://doi.org/10.3389/fmicb.2021.647183
10. Bukhari T, Pevsner R, Herren Jk. 2022. Microsporidia: a promising vector control tool for residual malaria transmission. Front Trop Dis 3:1–18. https://doi.org/10.3389/fitd.2022.957109
11. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, Li Y, Ye J, Yu C, Li Z, Zhang X, Wang J, Yang H, Fang L, Chen Q. 2018. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience 7:1–6. https://doi.org/10.1093/gigascience/gix120
12. Andrews S. 2010. FastQC a quality control tool for high throughput sequence data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
13. Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048. https://doi.org/10.1093/bioinformatics/btw354
14. Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61–5. https://doi.org/10.1093/nar/gkl842
15. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–45. https://doi.org/10.1093/nar/gkv1189
16. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324
17. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. Gigascience 10:1–4. https://doi.org/10.1093/gigascience/giab008
18. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol 20:257. https://doi.org/10.1186/s13059-019-1891-0
19. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595
20. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29:2607–2618. https://doi.org/10.1093/nar/29.12.2607
21. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA 117:9451–9457. https://doi.org/10.1073/pnas.1921046117
22. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351
23. Aurrecoechea C, Barreto A, Brestelli J, Brunk BP, Caler EV, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Iodice J, Kissinger JC, Kraemer ET, Li W, Nayak V, Pennington C, Pinney DF, Pitts B, Roos DS, Srinivasamoorthy G, Stoeckert CJ, Treatman C, Wang H. 2011. AmoebaDB and MicrosporidiaDB: functional genomic resources for *Amoebozoa* and *Microsporidia* species. Nucleic Acids Res 39:D612–9. https://doi.org/10.1093/nar/gkq1006
24. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086
25. Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:1–4. https://doi.org/10.1186/s13059-019-1832-y
26. Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650. https://doi.org/10.1093/molbev/msp077
27. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010
28. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:1–14. https://doi.org/10.1186/s13059-015-0721-2
29. Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol 61:1061–1067. https://doi.org/10.1093/sysbio/sys062