Near chromosome-level genome assembly of the microsporidium *Hamiltosporidium tvaerminnensis*

Running head: *Hamiltosporidium tvaerminnensis* genome report

Pascal Angst[1*], Jean-François Pombert[2], Dieter Ebert[1,†], and Peter D. Fields[1,†]

[1] Department of Environmental Sciences, Zoology, University of Basel, Basel, Switzerland
[2] Department of Biology, Illinois Institute of Technology, Chicago, Illinois, USA.
[*] Corresponding author
[†] Shared last author

Pascal Angst
Department of Environmental Sciences, Zoology
University of Basel
Vesalgasse 1, 4051 Basel, +4161 207 03 75, pascal.angst@unibas.ch

OrcIDs
Pascal Angst: 0000-0002-8654-2251
Jean-François Pombert: 0000-0002-3449-8375
Dieter Ebert: 0000-0003-2653-3772
Peter D. Fields: 0000-0003-2959-2524

Keywords: fungi, microsporidia, telomere-to-telomere, Iso-seq, alternative polyadenylation, splicing efficiency, *Hamiltosporidium tvaerminnensis*

Abstract

Microsporidia are intracellular parasitic fungi whose genomes rank among the smallest of all known eukaryotes. A number of outstanding questions remain concerning the evolution of their large-scale variation in genome architecture, responsible for genome size variation of more than an order of magnitude. This genome report presents a first near-chromosomal assembly of a large-genome microsporidium, *Hamiltosporidium tvaerminnensis*. Combined Oxford Nanopore, Pacific Biosciences, and Illumina sequencing led to a genome assembly of 17 contigs, 11 of which represent complete chromosomes. Our assembly is 21.64 Mb in length, has an N50 of 1.44 Mb, and consists of 39.56 % interspersed repeats. We introduce a novel approach in microsporidia, PacBio Iso-Seq, as part of a larger annotation pipeline for obtaining high-quality annotations of 3,573 protein-coding genes. Based on direct evidence from the full-length Iso-Seq transcripts, we present evidence for alternative polyadenylation and variation in splicing efficiency, which are potential regulation mechanisms for gene expression in microsporidia. The generated high-quality genome assembly is a necessary resource for comparative genomics that will help elucidate the evolution of genome architecture in response to intracellular parasitism.

Significance

Microsporidia are a model for genome evolution in response to intracellular parasitism, but we lack high-quality resources from species with large genomes. We present a near complete assembly of a large-genome microsporidium, *Hamiltosporidium tvaerminnensis*, and obtain high-quality gene annotations through full-length transcripts using Iso-Seq, a

52 novel approach in microsporidia. Our study provides insights into gene regulation and paves
53 the way for comparative genomic analyses aiming to understand the evolution of genome
54 reduction and expansion in these intracellular parasites.

55 Introduction

56

57 Microsporidia have become a valuable model in evolutionary biology for studying extreme
58 parasitism (Murareanu et al., 2021; Wadi & Reinke, 2020). Compared to their fungi relatives,
59 the genomes of microsporidia have been strongly shaped by their specialized life history of
60 obligate intracellular parasitism (Corradi, 2015). In particular, the genomes of most
61 microsporidia are very compact and reduced (Murareanu et al., 2021). However, genome
62 sizes of microsporidia vary by more than an order of magnitude, making them a suitable
63 model for studying genome evolution in eukaryotes and parasites (Wadi & Reinke, 2020). A
64 powerful approach for this is comparative genomics (i.e., comparing the presence or
65 absence of genetic sequences, their relative location, and their abundance among different
66 species) for which high-quality genome assemblies are a prerequisite (Kille et al., 2022).
67 Increasing interest in microsporidia has boosted available genome resources, but most
68 published genomes are drafts (Williams et al., 2022). Following from recent successful
69 efforts to obtain telomere-to-telomere assemblies in *Encephalitozoon*, a genus of
70 microsporidia with small genomes (Mascarenhas dos Santos et al., 2023), we focus here on
71 a large-genome species, *Hamiltosporidium tvaerminnensis*.

72

73 First noted by Green (1957) as *Octosporea bayeri*, the reclassified *H. tvaerminnensis* has
74 become a model in the study of host–parasite interactions with its only host *Daphnia*

magna, a planktonic microcrustacean (Altermatt et al., 2007; Lass & Ebert, 2005; Orlansky & Ben-Ami, 2019; Routtu & Ebert, 2015; Santos & Ebert, 2022; Vizoso & Ebert, 2004). Haag et al. (2011) provided the cytological and molecular description of *H. tvaerminnensis*, and Corradi et al. (2009) produced the first draft of its genome. Recently, genetic and genomic studies have started using *H. tvaerminnensis* to investigate the evolution of variation in microsporidian genome architecture (Angst et al., 2022, 2023; de Albuquerque et al., 2020; Haag et al., 2013, 2020). *Hamiltosporidium tvaerminnensis* has a remarkably large genome compared to other microsporidia, partly due to the proliferation of transposable genomic elements (de Albuquerque et al., 2020; Parisot et al., 2014). Draft genomes are available for the nearest known relative, *H. magnivora*, and more distantly related species, *Thelohania contejeani* (Cormier, Wattier, et al., 2021) and *Edhazardia aedis* (Desjardins et al., 2015).

Demands for comparative genomic approaches are increasing, calling for an improved genome of *H. tvaerminnensis*, a leading model organism among large-genome microsporidia. A high-quality genome will allow advancements in the study of genome architecture evolution in microsporidia; for example, through comparison with chromosomal assemblies of *Encephalitozoon*. We have integrated Pacific Biosciences (PacBio) Continuous Long Reads (CLR), Oxford Nanopore Technologies (ONT), and Illumina DNA sequencing for a near-chromosomal assembly, and we employed PacBio Iso-Seq for high-quality gene annotation. The first use of Iso-Seq for a microsporidium provides insights into microsporidian alternative polyadenylation (APA) and splicing efficiency as potential regulation mechanisms of gene expression in reduced genomes.

98 # Results and Discussion

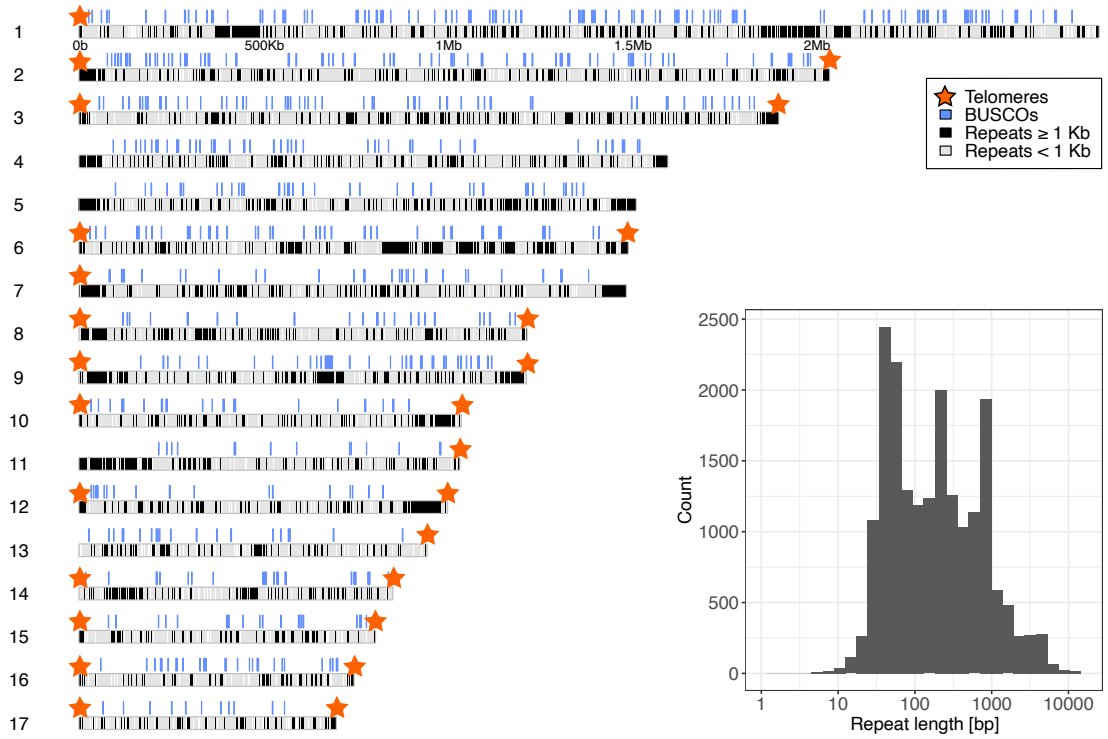
99

100 ## Genome Sequencing and Assembly

101 We generated a genome assembly of *H. tvaerminnensis* based on PacBio CLR and ONT
102 sequencing reads (total/N50: 17.53 Gb/1.54 Kb and 2.75 Gb/32.68 Kb, respectively). In a
103 subsequent step, we polished this assembly using Illumina paired-end (PE) sequencing reads
104 (total: 66 Gb). The final assembly consisted of 21.64 Mb in 17 contigs with an N50 of 1.44
105 Mb (Table 1, Figure 1). We used BUSCO to assess the biological completeness of our
106 assembly and found a BUSCO score of 94 %, a relatively high value for a large-genome
107 microsporidium (Cormier, Chebbi, et al., 2021). The structural completeness of our assembly
108 was assessed using scripts from (Mascarenhas dos Santos et al., 2023), which allow for the
109 identification of telomeres. Telomeres are identifiable by terminal, short, tandem repeats–
110 $(TTAGGG)_n$ in the case of *H. tvaerminnensis*, as is the case in most fungi (Rahnama et al.,
111 2020). Eleven contigs represented complete chromosomes flanked by telomeres at both
112 ends (Figure 1, Table S1), four contigs had only one telomere, and two had none. Therefore,
113 we estimated the genome contains 13 – 17 chromosomes. Large multicopy repeats arrayed
114 in tandem and scattered throughout the genome prevented us from fully assembling the
115 chromosomes with major repetitive loci abutting contig ends without telomeres (Figure 1).
116 Overall, 39.56 % of the genome consisted of interspersed repeats (retroelements 7.07 %,
117 DNA transposons 7.04 %, other 25.45 %). High numbers of repeats are common in large-
118 genome microsporidia (de Albuquerque et al., 2020), explaining why previous attempts to
119 assemble the genome of *H. tvaerminnensis* have been less successful in both completeness
120 and contiguity (size = 18.34 Mb; number of contigs = 2,915; N50 = 9.58 Kb; BUSCO score =

87.6 %) (Haag et al., 2020) compared to our long-read sequencing-based assembly. The highest quality genomes within the microsporidia are from the genus *Encephalitozoon*, which have an order of magnitude smaller genomes than *H. tvaerminnensis* and telomeric repeats of $(TTAGG)_n$ on both ends of all eleven chromosomes (Mascarenhas dos Santos et al., 2023).

Figure 1. Karyoplot of *H. tvaerminnensis*. Eleven of the 17 contigs have two telomeres, i.e., represent full chromosomes. In total, 39.56 % of the genome are interspersed repeats and low complexity DNA sequences. The inset shows the repeat length distribution on a base-10 logarithmic scale.



136 **Table 1. Statistics of the assembly and annotation of the *H. tvaerminnensis* genome.**

Assembly	
Total length (Mb)	21.64
GC content (%)	26.60
Contig N50 (Mb)	1.44
Contig number	17
BUSCO completeness score (%)	94
Annotation	
Total length of repeats (Mb)	8.56
Number of protein-coding genes	3,573
Mean gene length (bp)	1,391
Number of predicted introns	98
Number of Iso-Seq confirmed introns	17
Number of genes with confirmed introns	15
Mean intron length (bp)	28

137

138 ## Iso-Seq for Genome Annotation

139 RNA-seq data from intracellular parasites are sparse, and even more sparse are full-length
140 transcript sequences. Iso-Seq helped us to predict both gene structures and untranslated
141 regions (UTRs) more accurately than was previously done for microsporidia. Using 8.5 Gb of
142 Iso-Seq reads (so-called circular consensus sequence (CCS) of cDNA), we obtained
143 annotations for 3,573 protein-coding genes with a mean length of 1,391 bp (Table 1). The
144 number of genes in microsporidia ranges from about 2,000 to 4,000 and is positively
145 correlated with genome size (Jespersen et al., 2022). Compared to previous work, we
146 obtained fewer fragmented genes; the score of fragmented BUSCOs in our annotations was
147 1.8 % (previously 3.8 %; Haag et al., 2020). We also obtained many fewer introns (98) than
148 previously predicted (890), but our number is still higher than estimates for other large-
149 genome microsporidia annotated with the help of (short-read) RNA-seq (Desjardins et al.,
150 2015). After observing the discrepancy between annotations with and without RNA-seq
151 data, we suspected introns may not be efficiently spliced.

152

153 ## Splicing Efficiency Analysis

154 Species with reduced genomes, sparse introns or no introns, and reduced spliceosomes
155 have been suggested to exhibit low splicing efficiencies (Schärfen et al., 2022; Whelan et al.,
156 2019). Microsporidia are such species. For example, in *E. cuniculi*, splicing efficiency for
157 small introns is only around 20 % (Grisdale et al., 2013). We used direct evidence from full-
158 length transcripts and found high variation in splicing efficiency in *H. tvaerminnensis* as has
159 previously been reported for the reduced genomes of the red alga *Cyanidioschyzon merolae*
160 (Schärfen et al., 2022) and for parasitic fungi (Sieber et al., 2018). We used 17 introns from
161 15 genes with good representation in the Iso-Seq data (Table S2). These introns had a mean
162 length of 28 bp. The percentages of spliced transcripts ranged from 7.5 % to 92.3 % (Table
163 S2) with a mean of 54.53 %. We did not find a statistical correlation between percent spliced
164 and the length of introns. In conclusion, we found fewer introns in *H. tvaerminnensis* than
165 previously suggested and intermediate splicing efficiencies, which is in line with theorized
166 patterns for reduced-genome species. All but two introns identified in *H. tvaerminnensis*
167 featured the same splice sites as have been found in (non-)parasitic fungi (i.e., 5'GU...AG3')
168 (Kupfer et al., 2004). The other two introns had splice sites 5'AA…AG3' and 5'AU…GA3'.
169

170 ## Alternative Polyadenylation

171 The mechanisms of gene regulation in microsporidia are unexplored. APA is a widespread
172 and conserved regulatory mechanism by which the availability of multiple polyadenylation
173 sites can lead to variation in 3'-UTR lengths of mRNA isoforms and thus 3'-UTR content,
174 which strongly influences gene expression (Mitschka & Mayr, 2022). Here, Iso-Seq data
175 allowed us to investigate the occurrence of mRNA isoforms encoding the same protein but
176 differing in their 3'-UTRs resulting from APA. By observing multiple 3'-ends for many genes,

177 we found direct evidence for APA in the full-length transcripts of *H. tvaerminnensis*.
178 Upstream of the 3'-ends, a motif analysis of the 500 most expressed genes revealed
179 enrichment of an UAAA tetramer. The UAAA tetramer was found in 95.8 % of the genes and
180 87.7 % of them were part of an A[A/U]UAAA hexamer on average 20 nucleotides upstream
181 of the most common 3'-ends. A[A/U]UAAA is a known motif for cleavage of protein-coding
182 mRNAs in metazoa (Neve et al., 2017) and is shown to be enriched upstream of 3'-ends in
183 the bee-infecting microsporidium *Nosema ceranae* (Chen et al., 2021). These motifs were
184 slightly less enriched upstream of the second and third most common 3'-ends (Table S3). No
185 motif was strongly enriched that did not contain an UAAA tetramer.
186

187 Materials and Methods

188

189 Sample preparation and sequencing

190 The *H. tvaerminnensis* isolate used in this study was obtained from the *D. magna* genotype
191 FI-OER-3-3, collected in July 2005 from a rockpool (59°47'18.9"N 23°10'26.9"E) on the Island
192 Oeren in the Tvaerminne archipelago, Southwestern Finland. The same isolate has been
193 Illumina sequenced for earlier genome drafts (Corradi et al., 2009; Haag et al., 2020) and a
194 biogeographic study (Angst et al., 2022). Genomic Illumina reads from the latter were
195 reused here (NCBI database, SRA accession: SRX13146514, Bioproject ID: PRJNA780787).
196 We generated long-read sequencing data by applying a protocol described in Fields et al.,
197 (2015) for obtaining microbiota-free *D. magna* tissue, but retaining microsporidia, from
198 which we extracted high-molecular-weight DNA using Genomic-tips (QIAGEN). A standard
199 PacBio genomic DNA library and sequencing on a SMRTcell in CLR mode on a Sequel I have

200 previously been done at the Quantitative Genomics Facility service platform at the
201 Department of Biosystem Science and Engineering (D-BSSE, ETH) Basel, Switzerland (NCBI
202 database, SRA accession: SRR23724335, Bioproject ID: PRJNA778105). Additionally, a ONT
203 genomic DNA library was prepared with the SQK-LSK110 ligation kit and sequenced on a
204 MinION device with a Spot-ON Flow Cell (R9.4.1).
205
206 To extract full-length transcripts, we used the RNeasy extraction kit (QIAGEN)
207 supplying Proteinase K for protein digestion. The NEBNext Single Cell/Low Input cDNA
208 Synthesis and Amplification Module kit (New England Biolabs) was used to generate cDNA
209 from the extraction. An Iso-Seq cDNA library was generated with the Iso-Seq Express Oligo
210 Kit and the SMRTbell Express Template Prep Kit 2.0 (PacBio). The Iso-Seq library was
211 sequenced using a single SMRTcell on a Sequel I at the D-BSSE (Basel, Switzerland) and
212 processed with the *isoseq3* pipeline (https://github.com/PacificBiosciences/IsoSeq),
213 resulting in approximately 8.5 Gb of CCS data.
214
215 Genome assembly
216 An assembly based on the PacBio CLR sequencing data, generated with canu v.2.1.1 (Koren
217 et al., 2017), was extended *in-silico* using a seed-based approach with Consed v.29 (Gordon
218 et al., 1998) as described in Pombert et al. (2013). Independently, the Nanopore sequencing
219 reads were basecalled with bonito v.0.5.3 (https://github.com/nanoporetech/bonito) using
220 the dna_r9.4.1_e8.1_sup@v3.3 model and then assembled with nextDenovo v.2.5.0
221 (https://github.com/Nextomics/NextDenovo) using an expected genome size of 20 Mb. The
222 Nanopore nextDenovo assembly was polished with the Illumina reads and the PacBio CLR
223 reads using nextPolish v.1.4.1 (Hu et al., 2020), Medaka v.1.7.2

224 (https://github.com/nanoporetech/medaka), and two runs of Pilon v.1.24 (Walker et al.,
225 2014) and then merged with the extended PacBio CLR assembly using quickmerge v.0.3
226 (Solares et al., 2018) and manual curation. The absence of contaminants in this final
227 assembly was ascertained by BLAST homology searches against the NCBI database and its
228 completeness was assessed with check_for_telomeres.pl v.0.3 (Mascarenhas dos Santos et
229 al., 2023). Repeats in the assembly were detected with RepeatModeler v.2.0.2 including the
230 LTR pipeline (Flynn et al., 2020) and ReapeatMasker v.4.1.2 (Smit et al., 2013) and plotted
231 with karyoploteR v.1.24.0 (Gel & Serra, 2017) in R v.4.2.2 (R Core Team, 2022).
232

233 Genome annotation

234 The *H. tvaerminnensis* genome was annotated by feeding the Iso-Seq transcriptome data as
235 extrinsic evidence to the fungal annotation pipeline funannotate v.1.8.13 (Palmer & Stajich,
236 2022). Briefly, as per guideline for gene prediction using long RNA reads
237 (https://github.com/Gaius-
238 Augustus/BRAKER/blob/master/docs/long_reads/long_read_protocol.md), we mapped the
239 whole-transcript reads from the CCS FASTA produced by *isoseq3* to the repeat-masked
240 genome using Minimap2 v.2.22 (Li, 2018) with the flags '-ax splice –secondary=no -C5'.
241 Approximately 25.5 % of the reads aligned to the parasite genome. We then collapsed the
242 total read set into distinct transcripts (while retaining isoforms) using the
243 *collapse_isoforms_by_sam.py* script provided by the *cDNA_Cupcake*
244 (https://github.com/Magdoll/cDNA_Cupcake) pipeline and ran GeneMarkS-T v.5.1 (Tang et
245 al., 2015) to predict protein-coding regions in the transcripts. The resultant FASTA and GFF
246 files were used for gene prediction and annotation with the wrapper tool funannotate
247 v.1.8.13 (Palmer & Stajich, 2022), which relies on TANTAN v.26 (Frith, 2011) for repeat-

248 masking. Funannotate operated AUGUSTUS v.3.4.0 (Stanke et al., 2006, 2008), GeneMark-ES
249 v.4.62 (Brůna et al., 2020), GlimmerHMM v.3.0.4 (Majoros et al., 2004), SNAP v.2013_11_29
250 (Korf, 2004), and EVidenceModeler v.1.1.1 (Haas et al., 2008) for ab-initio gene prediction
251 and consensus gene structure generation. Functional annotation in funannoate relies on the
252 combined evidence from InterProScan v.5.55_88.0 (Jones et al., 2014), eggNOG-mapper
253 v.2.1.9 (Cantalapiedra et al., 2021; Huerta-Cepas et al., 2019), Phobius v.1.01 (Käll et al.,
254 2004), SignalP v.5.0b (Almagro Armenteros et al., 2019), and antiSMASH v.6.0 (Blin et al.,
255 2021) as well as on comparisons to Pfam (Mistry et al., 2021), UniProt (The UniProt
256 Consortium, 2022), MEROPS (Rawlings et al., 2014), dbCAN (Yin et al., 2012) and BUSCO
257 v.5.4.3 (Manni et al., 2021) databases. Also, the completeness was assessed using BUSCO
258 with its microsporidia_odb10 database (Creation date: 2020-08-05). For proteins without
259 functional annotation from this pipeline, we lifted-over functional annotations from the
260 species reference proteome (UniProt database, Proteome ID: UP000292282) using BLAST
261 v.2.3.0+ (Camacho et al., 2009), AGAT v.1.0.0 (Dainat, 2022), and GFF3sort v.1.0.0 (Zhu et
262 al., 2017).
263

264 ## Full-length transcript analyses

265 For the analysis of splicing efficiency and APA, we followed the methodology described in
266 Schärfen et al. (2022). Briefly, starting from the BAM file, which was provided by the
267 sequencing facility, we used lima v.2.6.0 (https://github.com/PacificBiosciences/barcoding)
268 in "-isoseq" mode to remove IsoSeq template-switching oligo sequences and to orient
269 subreads in the correct 5' to 3' direction. After aligning subreads to our genome using
270 minimap2 in "splice:hq" mode, we used SAMtools v.1.16.1 (Li et al., 2009) and BEDtools
271 v.2.30.0 (Quinlan, 2014) to identify intron-spanning reads. The percentage of spliced introns

272 was calculated by dividing the number of spliced reads by the total number of reads for
273 each intron times 100. For this analysis, introns were manually selected based on Iso-Seq
274 coverage level. A potential correlation between the percent of spliced reads and intron
275 length was tested in R.
276
277 For the analysis of APA, briefly, we used SAMtools, BEDtools, and pysam v.0.20.0
278 (https://github.com/pysam-developers/pysam) to extract each read's 3'-end coordinate
279 following Schärfen et al. (2022). Then, we drew the distribution of 3'-end coordinates as a
280 histogram separately for each gene and applied the find_peaks algorithm from scipy.signal
281 v.1.7.3 (Virtanen et al., 2020) with height equal to four and distance equal to 18. Hence, we
282 found the number of peaks (=3'-end coordinates) per gene. We then separately extracted
283 sequences of the 500 most expressed genes up to 40 nucleotides upstream of the highest
284 peak, the second highest peak, and the third highest peak using seqtk v.1.3-r106
285 (https://github.com/lh3/seqtk). For each set of sequences, we did a motif analysis using
286 XSTREME v5.5.0 from the MEME Suite (Grant & Bailey, 2021) with a minimum motif length
287 of four and determined the distance of the obtained motif from the peak.

288 Data Availability Statement

289 Raw data is deposited at the NCBI SRA database and the assembled genome as well as
290 predicted set of protein sequences are available at NCBI GenBank (FIOER33 v3; BioProject ID
291 PRJNA778105) and at https://doi.org.10.6084/m9.figshare.22794887 and will be publicly
292 available upon acceptance.
293
294

295 ## Conflict of Interest

296 The authors have no conflicts of interest to declare.

297 ## Author Contributions

298 All authors designed the study. DE collected the sample. PA and PDF performed the
299 molecular work. PA, JFP, and PDF analyzed the data. PA wrote the manuscript. All authors
300 reviewed the manuscript.

301 ## Acknowledgements

302 We thank Jürgen Hottinger, Urs Stiefel, Michelle Krebs, and Alix Thivolle for help in the
303 laboratory. We thank members of the Ebert group for providing feedback on the study and
304 the manuscript.

305 ## Funding

306 This work was supported by the Swiss National Science Foundation (SNSF) (grant numbers
307 310030B_166677 and 310030_188887 to DE).

308

309 ## References

310 Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O.,
311 Brunak, S., von Heijne, G., & Nielsen, H. (2019). SignalP 5.0 improves signal peptide
312 predictions using deep neural networks. *Nature Biotechnology*, *37*(4), Article 4.
313 https://doi.org/10.1038/s41587-019-0036-z

314 Altermatt, F., Hottinger, J., & Ebert, D. (2007). Parasites promote host gene flow in a
315 metapopulation. *Evolutionary Ecology*, *21*(4), 561–575.
316 https://doi.org/10.1007/s10682-006-9136-6
317 Angst, P., Ebert, D., & Fields, P. D. (2022). Demographic history shapes genomic variation in
318 an intracellular parasite with a wide geographical distribution. *Molecular Ecology*,
319 *31*(9), 2528–2544. https://doi.org/10.1111/mec.16419
320 Angst, P., Ebert, D., & Fields, P. D. (2023). Population genetic analysis of the microsporidium
321 *Ordospora colligata* reveals the role of natural selection and phylogeography on its
322 extremely compact and reduced genome. *G3 Genes|Genomes|Genetics*, jkad017.
323 https://doi.org/10.1093/g3journal/jkad017
324 Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., van Wezel, G. P., Medema, M. H.,
325 & Weber, T. (2021). antiSMASH 6.0: Improving cluster detection and comparison
326 capabilities. *Nucleic Acids Research*, *49*(W1), W29–W35.
327 https://doi.org/10.1093/nar/gkab335
328 Brůna, T., Lomsadze, A., & Borodovsky, M. (2020). GeneMark-EP+: Eukaryotic gene
329 prediction with self-training in the space of genes and proteins. *NAR Genomics and*
330 *Bioinformatics*, *2*(2), lqaa026. https://doi.org/10.1093/nargab/lqaa026
331 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T.
332 L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421.
333 https://doi.org/10.1186/1471-2105-10-421
334 Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021).
335 eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain
336 Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, *38*(12),
337 5825–5829. https://doi.org/10.1093/molbev/msab293

338 Chen, H.-Z., Fan, X.-X., Fan, Y.-C., Wang, J., Zhu, Z.-W., Jiang, H.-B., Zhang, W.-D., Long, Q.,
339 Xiong, C.-L., Zheng, Y.-Z., Fu, Z.-M., Xu, G.-J., Chen, D.-F., & Guo, R. (2021). Analysis of
340 alternative splicing and alternative polyadenylation of *Nosema ceranae* genes.
341 *Mycosystema*, *40*(1), 161–173. https://doi.org/10.13346/j.mycosystema.200216
342 Cormier, A., Chebbi, M. A., Giraud, I., Wattier, R., Teixeira, M., Gilbert, C., Rigaud, T., &
343 Cordaux, R. (2021). Comparative Genomics of Strictly Vertically Transmitted,
344 Feminizing Microsporidia Endosymbionts of Amphipod Crustaceans. *Genome Biology*
345 *and Evolution*, *13*(1), evaa245. https://doi.org/10.1093/gbe/evaa245
346 Cormier, A., Wattier, R., Giraud, I., Teixeira, M., Grandjean, F., Rigaud, T., & Cordaux, R.
347 (2021). Draft Genome Sequences of *Thelohania contejeani* and *Cucumispora*
348 *dikerogammari*, Pathogenic Microsporidia of Freshwater Crustaceans. *Microbiology*
349 *Resource Announcements*, *10*(2), e01346-20. https://doi.org/10.1128/MRA.01346-20
350 Corradi, N. (2015). Microsporidia: Eukaryotic Intracellular Parasites Shaped by Gene Loss
351 and Horizontal Gene Transfers. *Annual Review of Microbiology*, *69*(1), 167–183.
352 https://doi.org/10.1146/annurev-micro-091014-104136
353 Corradi, N., Haag, K. L., Pombert, J.-F., Ebert, D., & Keeling, P. J. (2009). Draft genome
354 sequence of the *Daphnia* pathogen *Octosporea bayeri*: Insights into the gene content
355 of a large microsporidian genome and a model for host-parasite interactions.
356 *Genome Biology*, *10*(10), R106. https://doi.org/10.1186/gb-2009-10-10-r106
357 Dainat, J. (2022). *AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF*
358 *format.* Zenodo. https://doi.org/10.5281/zenodo.7255559
359 de Albuquerque, N. R. M., Ebert, D., & Haag, K. L. (2020). Transposable element abundance
360 correlates with mode of transmission in microsporidian parasites. *Mobile DNA*, *11*,
361 19. https://doi.org/10.1186/s13100-020-00218-8

362 Desjardins, C. A., Sanscrainte, N. D., Goldberg, J. M., Heiman, D., Young, S., Zeng, Q.,
363 Madhani, H. D., Becnel, J. J., & Cuomo, C. A. (2015). Contrasting host–pathogen
364 interactions and genome evolution in two generalist and specialist microsporidian
365 pathogens of mosquitoes. *Nature Communications*, *6*, 7121.
366 https://doi.org/10.1038/ncomms8121
367 Fields, P. D., Reisser, C., Dukić, M., Haag, C. R., & Ebert, D. (2015). Genes mirror geography in
368 *Daphnia magna*. *Molecular Ecology*, *24*(17), 4521–4536.
369 https://doi.org/10.1111/mec.13324
370 Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020).
371 RepeatModeler2 for automated genomic discovery of transposable element families.
372 *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457.
373 https://doi.org/10.1073/pnas.1921046117
374 Frith, M. C. (2011). A new repeat-masking method enables specific detection of homologous
375 sequences. *Nucleic Acids Research*, *39*(4), e23. https://doi.org/10.1093/nar/gkq1212
376 Gel, B., & Serra, E. (2017). karyoploteR: An R/Bioconductor package to plot customizable
377 genomes displaying arbitrary data. *Bioinformatics*, *33*(19), 3088–3090.
378 https://doi.org/10.1093/bioinformatics/btx346
379 Gordon, D., Abajian, C., & Green, P. (1998). Consed: A Graphical Tool for Sequence Finishing.
380 *Genome Research*, *8*(3), 195–202. https://doi.org/10.1101/gr.8.3.195
381 Grant, C. E., & Bailey, T. L. (2021). *XSTREME: Comprehensive motif analysis of biological*
382 *sequence datasets* (p. 2021.09.02.458722). bioRxiv.
383 https://doi.org/10.1101/2021.09.02.458722
384 Green, J. (1957). Parasites and epibionts of Cladocera in rock pools of Tvärminne
385 archipelago. *Archivum Societatis Zoologicae Botanicae Fennicae 'Vanamo*, *12*, 5–12.

386 Grisdale, C. J., Bowers, L. C., Didier, E. S., & Fast, N. M. (2013). Transcriptome analysis of the
387 parasite *Encephalitozoon cuniculi*: An in-depth examination of pre-mRNA splicing in a
388 reduced eukaryote. *BMC Genomics*, *14*(1), 207. https://doi.org/10.1186/1471-2164-
389 14-207
390 Haag, K. L., Larsson, J. I. R., Refardt, D., & Ebert, D. (2011). Cytological and molecular
391 description of *Hamiltosporidium tvaerminnensis* gen. Et sp. Nov., a microsporidian
392 parasite of *Daphnia magna*, and establishment of *Hamiltosporidium magnivora*
393 comb. Nov. *Parasitology*, *138*(4), 447–462.
394 https://doi.org/10.1017/S0031182010001393
395 Haag, K. L., Pombert, J.-F., Sun, Y., de Albuquerque, N. R. M., Batliner, B., Fields, P., Lopes, T.
396 F., & Ebert, D. (2020). Microsporidia with Vertical Transmission Were Likely Shaped
397 by Nonadaptive Processes. *Genome Biology and Evolution*, *12*(1), 3599–3614.
398 https://doi.org/10.1093/gbe/evz270
399 Haag, K. L., Traunecker, E., & Ebert, D. (2013). Single-nucleotide polymorphisms of two
400 closely related microsporidian parasites suggest a clonal population expansion after
401 the last glaciation. *Molecular Ecology*, *22*(2), 314–326.
402 https://doi.org/10.1111/mec.12126
403 Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., &
404 Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using
405 EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome*
406 *Biology*, *9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7
407 Hu, J., Fan, J., Sun, Z., & Liu, S. (2020). NextPolish: A fast and efficient genome polishing tool
408 for long-read assembly. *Bioinformatics*, *36*(7), 2253–2255.
409 https://doi.org/10.1093/bioinformatics/btz891

410 Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H.,
411 Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019).
412 eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology
413 resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1),
414 D309–D314. https://doi.org/10.1093/nar/gky1085
415 Jespersen, N., Monrroy, L., & Barandun, J. (2022). Impact of Genome Reduction in
416 Microsporidia. In L. M. Weiss & A. W. Reinke (Eds.), *Microsporidia: Current Advances*
417 *in Biology* (pp. 1–42). Springer International Publishing. https://doi.org/10.1007/978-
418 3-030-93306-7_1
419 Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J.,
420 Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew,
421 M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein
422 function classification. *Bioinformatics*, *30*(9), 1236–1240.
423 https://doi.org/10.1093/bioinformatics/btu031
424 Käll, L., Krogh, A., & Sonnhammer, E. L. L. (2004). A Combined Transmembrane Topology
425 and Signal Peptide Prediction Method. *Journal of Molecular Biology*, *338*(5), 1027–
426 1036. https://doi.org/10.1016/j.jmb.2004.03.016
427 Kille, B., Balaji, A., Sedlazeck, F. J., Nute, M., & Treangen, T. J. (2022). Multiple genome
428 alignment in the telomere-to-telomere assembly era. *Genome Biology*, *23*(1), 182.
429 https://doi.org/10.1186/s13059-022-02735-6
430 Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017).
431 Canu: Scalable and accurate long-read assembly via adaptive *k*-mer weighting and
432 repeat separation. *Genome Research*, *27*(5), 722–736.
433 https://doi.org/10.1101/gr.215087.116

434 Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*(1), 59.
435 https://doi.org/10.1186/1471-2105-5-59
436 Kupfer, D. M., Drabenstot, S. D., Buchanan, K. L., Lai, H., Zhu, H., Dyer, D. W., Roe, B. A., &
437 Murphy, J. W. (2004). Introns and Splicing Elements of Five Diverse Fungi. *Eukaryotic*
438 *Cell*, *3*(5), 1088–1100. https://doi.org/10.1128/EC.3.5.1088-1100.2004
439 Lass, S., & Ebert, D. (2005). Apparent seasonality of parasite dynamics: Analysis of cyclic
440 prevalence patterns. *Proceedings of the Royal Society B: Biological Sciences*,
441 *273*(1583), 199–206. https://doi.org/10.1098/rspb.2005.3310
442 Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*,
443 *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191
444 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., &
445 Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools.
446 *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
447 Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: Two open
448 source ab initio eukaryotic gene-finders. *Bioinformatics*, *20*(16), 2878–2879.
449 https://doi.org/10.1093/bioinformatics/bth315
450 Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO
451 Update: Novel and Streamlined Workflows along with Broader and Deeper
452 Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.
453 *Molecular Biology and Evolution*, *38*(10), 4647–4654.
454 https://doi.org/10.1093/molbev/msab199
455 Mascarenhas dos Santos, A. C., Julian, A. T., Liang, P., Juárez, O., & Pombert, J.-F. (2023).
456 Telomere-to-Telomere genome assemblies of human-infecting *Encephalitozoon*
457 species. *BMC Genomics*, *24*(1), 237. https://doi.org/10.1186/s12864-023-09331-3

458 Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L.,
459 Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A.
460 (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1),
461 D412–D419. https://doi.org/10.1093/nar/gkaa913
462 Mitschka, S., & Mayr, C. (2022). Context-specific regulation and function of mRNA
463 alternative polyadenylation. *Nature Reviews Molecular Cell Biology*, *23*(12), Article
464 12. https://doi.org/10.1038/s41580-022-00507-5
465 Murareanu, B. M., Sukhdeo, R., Qu, R., Jiang, J., & Reinke, A. W. (2021). Generation of a
466 Microsporidia Species Attribute Database and Analysis of the Extensive Ecological
467 and Phenotypic Diversity of Microsporidia. *MBio*, *12*(3), e01490-21.
468 https://doi.org/10.1128/mBio.01490-21
469 Neve, J., Patel, R., Wang, Z., Louey, A., & Furger, A. M. (2017). Cleavage and
470 polyadenylation: Ending the message expands gene regulation. *RNA Biology*, *14*(7),
471 865–890. https://doi.org/10.1080/15476286.2017.1306171
472 Orlansky, S., & Ben-Ami, F. (2019). Genetic resistance and specificity in sister taxa of
473 *Daphnia*: Insights from the range of host susceptibilities. *Parasites & Vectors*, *12*(1),
474 545. https://doi.org/10.1186/s13071-019-3795-y
475 Palmer, J., & Stajich, J. (2022). *nextgenusfs/funannotate: Funannotate* (1.8.11). Zenodo.
476 https://zenodo.org/record/2604804
477 Parisot, N., Pelin, A., Gasc, C., Polonais, V., Belkorchia, A., Panek, J., El Alaoui, H., Biron, D. G.,
478 Brasset, E., Vaury, C., Peyret, P., Corradi, N., Peyretaillade, É., & Lerat, E. (2014).
479 Microsporidian genomes harbor a diverse array of transposable elements that
480 demonstrate an ancestry of horizontal exchange with metazoans. *Genome Biology
481 and Evolution*, *6*(9), 2289–2300. https://doi.org/10.1093/gbe/evu178

Pombert, J.-F., Xu, J., Smith, D. R., Heiman, D., Young, S., Cuomo, C. A., Weiss, L. M., & Keeling, P. J. (2013). Complete Genome Sequences from Three Genetically Distinct Strains Reveal High Intraspecies Genetic Diversity in the Microsporidian *Encephalitozoon cuniculi*. *Eukaryotic Cell*, *12*(4), 503–511. https://doi.org/10.1128/EC.00312-12

Quinlan, A. R. (2014). BEDTools: The Swiss-army tool for genome feature analysis. *Current Protocols in Bioinformatics*, *47*, 11.12.1-11.12.34. https://doi.org/10.1002/0471250953.bi1112s47

R Core Team. (2022). *R: The R Project for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rahnama, M., Novikova, O., Starnes, J. H., Zhang, S., Chen, L., & Farman, M. L. (2020). Transposon-mediated telomere destabilization: A driver of genome evolution in the blast fungus. *Nucleic Acids Research*, *48*(13), 7197–7217. https://doi.org/10.1093/nar/gkaa287

Rawlings, N. D., Waller, M., Barrett, A. J., & Bateman, A. (2014). MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, *42*(D1), D503–D509. https://doi.org/10.1093/nar/gkt953

Routtu, J., & Ebert, D. (2015). Genetic architecture of resistance in *Daphnia* hosts against two species of host-specific parasites. *Heredity*, *114*(2), 241–248. https://doi.org/10.1038/hdy.2014.97

Santos, J. L., & Ebert, D. (2022). The effects of temperature and host-parasite interactions on parasite persistence in a planktonic crustacean. *Journal of Freshwater Ecology*, *37*(1), 555–568. https://doi.org/10.1080/02705060.2022.2134219

505 Schärfen, L., Zigackova, D., Reimer, K. A., Stark, M. R., Slat, V. A., Francoeur, N. J., Wells, M.
506 L., Zhou, L., Blackshear, P. J., Neugebauer, K. M., & Rader, S. D. (2022). Identification
507 of Alternative Polyadenylation in *Cyanidioschyzon merolae* Through Long-Read
508 Sequencing of mRNA. *Frontiers in Genetics*, *12*.
509 https://www.frontiersin.org/articles/10.3389/fgene.2021.818697
510 Sieber, P., Voigt, K., Kämmer, P., Brunke, S., Schuster, S., & Linde, J. (2018). Comparative
511 Study on Alternative Splicing in Human Fungal Pathogens Suggests Its Involvement
512 During Host Invasion. *Frontiers in Microbiology*, *9*.
513 https://www.frontiersin.org/articles/10.3389/fmicb.2018.02313
514 Smit, A., Hubley, R., & Green, P. (2013). *RepeatMasker Open-4.0*.
515 http://www.repeatmasker.org
516 Solares, E. A., Chakraborty, M., Miller, D. E., Kalsow, S., Hall, K., Perera, A. G., Emerson, J. J.,
517 & Hawley, R. S. (2018). Rapid Low-Cost Assembly of the *Drosophila melanogaster*
518 Reference Genome Using Low-Coverage, Long-Read Sequencing. *G3*
519 *Genes|Genomes|Genetics*, *8*(10), 3143–3154.
520 https://doi.org/10.1534/g3.118.200162
521 Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically
522 mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*(5),
523 637–644. https://doi.org/10.1093/bioinformatics/btn013
524 Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in
525 eukaryotes with a generalized hidden Markov model that uses hints from external
526 sources. *BMC Bioinformatics*, *7*(1), 62. https://doi.org/10.1186/1471-2105-7-62

527 Tang, S., Lomsadze, A., & Borodovsky, M. (2015). Identification of protein coding regions in
528 RNA transcripts. *Nucleic Acids Research*, *43*(12), e78.
529 https://doi.org/10.1093/nar/gkv227
530 The UniProt Consortium. (2022). UniProt: The Universal Protein Knowledgebase in 2023.
531 *Nucleic Acids Research*, gkac1052. https://doi.org/10.1093/nar/gkac1052
532 Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D.,
533 Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M.,
534 Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E.,
535 … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific
536 computing in Python. *Nature Methods*, *17*(3), Article 3.
537 https://doi.org/10.1038/s41592-019-0686-2
538 Vizoso, D. B., & Ebert, D. (2004). Within-host dynamics of a microsporidium with horizontal
539 and vertical transmission: *Octosporea bayeri* in *Daphnia magna*. *Parasitology*, *128*(Pt
540 1), 31–38. https://doi.org/10.1017/s0031182003004293
541 Wadi, L., & Reinke, A. W. (2020). Evolution of microsporidia: An extremely successful group
542 of eukaryotic intracellular parasites. *PLOS Pathogens*, *16*(2), e1008276.
543 https://doi.org/10.1371/journal.ppat.1008276
544 Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A.,
545 Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for
546 Comprehensive Microbial Variant Detection and Genome Assembly Improvement.
547 *PLOS ONE*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963
548 Whelan, T. A., Lee, N. T., Lee, R. C. H., & Fast, N. M. (2019). Microsporidian Introns Retained
549 against a Background of Genome Reduction: Characterization of an Unusual Set of

Introns. *Genome Biology and Evolution*, *11*(1), 263–269. https://doi.org/10.1093/gbe/evy260

Williams, B. A. P., Williams, T. A., & Trew, J. (2022). Comparative Genomics of Microsporidia. In L. M. Weiss & A. W. Reinke (Eds.), *Microsporidia: Current Advances in Biology* (pp. 43–69). Springer International Publishing. https://doi.org/10.1007/978-3-030-93306-7_2

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., & Xu, Y. (2012). dbCAN: A web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, *40*(W1), W445–W451. https://doi.org/10.1093/nar/gks479

Zhu, T., Liang, C., Meng, Z., Guo, S., & Zhang, R. (2017). GFF3sort: A novel tool to sort GFF3 files for tabix indexing. *BMC Bioinformatics*, *18*(1), 482. https://doi.org/10.1186/s12859-017-1930-3