Variation in genome architecture and epigenetic modification across the microsporidia phylogeny

Running head: Microsporidia genetic and epigenetic architecture

Pascal Angst[1,*], Dieter Ebert[1], and Peter D. Fields[1]

[1] Department of Environmental Sciences, Zoology, University of Basel, Basel, Switzerland

[*] Corresponding 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
Dieter Ebert: 0000-0003-2653-3772
Peter D. Fields: 0000-0003-2959-2524

Keywords: methylation, epigenetics, genome architecture, synteny, microsporidia, *Daphnia*

Abstract

Microsporidia are a model clade for studying intracellular parasitism, being well-known for their streamlined genomes and their extreme life history. Although microsporidia are highly diverse and ecologically important to a broad range of hosts, previous research on genome architecture has focused primarily on the mammal-infecting genus *Encephalitozoon*. Here, we expand that work, testing the universality of the patterns observed in *Encephalitozoon* by investigating and comparing variation in genetic and epigenetic architectures in the high-quality genome assemblies of several major microsporidia clades. Our comparison of nine genomes, including the first genome assemblies of *Binucleata daphniae*, *Gurleya vavrai*, and *Conglomerata obtusa*, and revised, improved assemblies of *Glugoides intestinalis* and *Ordospora colligata*, found limited conservation of genetic and epigenetic architecture across all microsporidia, although many genomic characteristics, such as nucleotide composition and repeat

content, were shared between genomes of the same or related clades. For example, rRNA genes were hypermethylated in all species, but their position close to chromosome ends was only found in the *Encephalitozoon* and its sister clade. GC-content varied widely, linked to genome size, phylogenetic position and activity of repeat elements. These findings enhance our insight into genome evolution and, consistent with findings from other systems, suggest epigenetic modification as a regulatory mechanism of gene expression and repeat element activity in microsporidia. Our comparative genome analysis reveals higher variation among microsporidia than previously supposed.

Introduction

The study of parasite evolution looks at the mechanisms by which parasites adapt to exploit their hosts (Poulin & Randhawa, 2015; Schmid-Hempel, 2021). These adaptations include physical structures used to invade or attach to the host, secretory molecules that subvert the host's immune system or alter its behavior to the parasite's advantage, and genomic rearrangements that optimize gene transcription and genome replication. Genome evolution particularly has received much recent attention in the field of microsporidia. These fungi-related intracellular parasites exhibit high variation in genome length, possibly related to their life history (Haag et al., 2020), but although comparative studies of microsporidia have provided initial insights into what is arguably the most extreme form of parasitism, including the molecular basis of parasitic adaptations (Nakjang et al., 2013; Wadi & Reinke, 2020; Williams et al., 2022), the lack of high-quality genomes for most major microsporidia clades in this little understood branch of the tree of life has limited comparative studies of genome architecture to single genera (see e.g., Khalaf et al. (2024), Mascarenhas dos Santos et al. (2023)). The available short-read-based genome assemblies have provided inadequate insights into microsporidia genomic architecture, despite hinting at its high diversity.

Microsporidia exhibit highly derived genomic features for parasitism, including mitochondrion genome loss and nuclear genome compaction (Jespersen et al., 2022), the latter of which is highly variable among microsporidia clades, with different forms and levels of (non-)coding DNA loss. Extreme examples can be seen in the mammalian parasite genus *Enterocytozoon*, which lacks most genes involved in glycolysis and is therefore energetically completely dependent on its host (Wiredu Boakye et al., 2017), and in the Encephalitozoonidae family, which has reduced inter- and intragenic regions, including repeat content (Corradi & Slamovits, 2011; Mascarenhas dos Santos et al., 2023). An opposite phenomenon, genome expansion, is observed secondarily in microsporidia with mixed-mode (horizontal and vertical) transmission, such as *Nosema bombycis* and *Hamiltosporidium tvaerminnensis* (de Albuquerque et al., 2020; Parisot et al., 2014). This likely occurs because vertical transmission reduces effective population size and thus the efficacy of natural selection to limit genome expansion, e.g., through the proliferation of transposable elements (Haag et al., 2020). While the quality of microsporidian genome assemblies is not as essential when it comes to comparing the

length or presence of genetic elements, we do need high-quality assemblies to study the abundance and relative location of these genetic elements across species.

Although microsporidia are found in a wide range of host species, including humans and agriculturally important animals, from different environments, most microsporidia seem to occur in aquatic environments with the greatest richness found in Crustacea (Bojko & Stentiford, 2022). Indeed, thanks, perhaps, to the vast abundance of Crustacea and the opportunistic nature of microsporidia (Weiss & Becnel, 2014), Crustacea host microsporidia from all major clades (Bojko et al., 2022). Additionally, because many microcrustaceans such as the model system *Daphnia* are filter feeders, they accumulate infectious pathogen cells during feeding, which provides the initial contact point and opportunity for microsporidia–host interactions. *Daphnia* alone are known to be regularly infected by species from at least four of the seven major microsporidia clades, as well as an expanded microsporidium that still has a mitochondrion (Ebert, 2008; Haag et al., 2014).

In this comparative genomic study, we aim to better understand genomic evolution in microsporidia by elucidating the causes of variation in their genomic features, and the arrangement and epigenetic modifications of their genome. Only recently, and only for a few microsporidia genera, have genomic architecture and methylation patterns been described. For example, a recently-described genomic architecture for three species in the human-infecting genus *Encephalitozoon* found that each chromosome in the genomes of *Encephalitozoon* spp., from the ends to the core, consists of 5-mer telomeric repeats, telomere-associated repeat elements (TAREs), hypermethylated ribosomal RNA (rRNA) genes, less methylated subtelomeres, and a hypomethylated chromosome core (Mascarenhas dos Santos et al., 2023). Our aim is to evaluate the generality of this architecture by using high-quality genome resources to examine microsporidia from diverse clades of the phylum. We look at *Encephalitozoon intestinalis*, its close relative *Ordospora colligata*, *Vairimorpha necatrix* also of the Nosematida clade, *Glugoides intestinalis* of the sister clade Enterocytozoonida, *Hamiltosporidium tvaerminnensis* of the orphan clade, three microsporidia of the Amblyosporida clade (*Binucleata daphniae*, *Gurleya vavrai*, and *Conglomerata obtusa* (≡ *Larssonia obtusa*)), and one expanded microsporidium, *Mitosporidium daphniae*. We further explore the processes that drive genome rearrangements and epigenetic modifications in these microsporidia. Previous studies have shown a relationship between mode of transmission and repeat content in the microsporidia genome (de Albuquerque et al., 2020; Haag et al., 2020), suggesting that, for example, life history contributes to the evolution of genomic architecture. As part of our study on genome organization and (epi)genetic features in these diverse microsporidia, we also contribute here the first high-quality genome assemblies and epigenetic annotations for a number of understudied microsporidia clades.

Results

Genome assemblies

We generated genome assemblies for five microsporidia (Table 1), including first-time assemblies for *B. daphniae*, *G. vavrai*, and *C. obtusa*, as well as enhanced assemblies for *O. colligata* and *G. intestinalis*. To expand our taxon sampling, we also included available high-quality genomes of four other microsporidia: *E. intestinalis* (Mascarenhas dos Santos et al., 2023), *H. tvaerminnensis* (Angst et al., 2023), *M. daphniae* (Angst et al., in prep.), and *V. necatrix* (Svedberg et al., 2024) (Table 1). The overall length of the nine genome assemblies correlated negatively with their GC-content ($F(1,7) = 19.44$, $P = 0.003$, $R^2 = 0.74$; Figure 1), following the trend seen in all (draft) microsporidia reference genome assemblies available in the NCBI database (Figure S1, Table S1). The Amblyosporida, *H. tvaerminnensis*, and *V. necatrix* genomes had about 15 % lower GC-content (22 – 29 %) and were about three times longer (10 – 22 Mb) than the smaller *E. intestinalis*, *O. colligata*, *G. intestinalis*, and *M. daphniae* genomes (34 – 43 % and 3 – 6 Mb). The larger genomes had a higher repeat content (26 – 50 %) and a higher number of protein-coding genes (2,847 – 4,231 bp) than the smaller genomes (12 – 20 % and 1,768 – 2,840 bp), except for *G. intestinalis*, whose genome consisted of 43 % repeats (Positive correlation between assembly and repeat length: $F(1,7) = 36.35$, $P = 0.001$, $R^2 = 0.84$). The protein-coding genes in larger genomes tended to be shorter on average (952 – 1,081 bp) than in shorter genomes (1,019 – 1,447 bp), except for *H. tvaerminnensis* (1,391 bp) whose protein-coding genes were closest to the expanded microsporidium *M. daphniae* (1,447 bp) in average length.

Table 1: Assembly information for the analyzed microsporidia genomes. BUSCO completeness was calculated using the microsporidia specific microsporidia_odb10 database, which is not specific to expanded microsporidia like *M. daphniae*.

Glugoides intestinalis	***Gurleya vavrai***	***Hamiltosporidium tvaerminnensis***	***Mitosporidium daphniae***	***Ordospora colligata***	***Vairimorpha necatrix***
5.13	16.37	21.64	5.83	3.18	15.01
34.06	21.97	26.6	42.68	37.53	28.34
439,321	989,478	1,444,059	864,371	374,328	1,281,651
13	20	17	10	9	12
83.8	90.3	94	37.6	99	96.8
2.21 43.06	4.63 28.27	8.56 39.56	0.69 11.79	0.62 19.54	7.45 49.68
2,331	3,537	3,573	2,840	1,768	2,971
1,127	978	1,391	1,447	1,114	1,081
PacBio HiFi	ONT plus Illumina	PacBio CLR plus ONT	ONT plus Illumina	PacBio HiFi	PacBio HiFi
new	new	Angst et al. (2023)	Angst et al. (in prep.)	new	Svedberg et al. (2024)

155

	Binucleata daphniae	*Conglomerata obtusa*	*Encephalitozoon intestinalis*
Total length (Mb)	9.58	18.13	2.61
GC-content (%)	29.04	25.59	41.94
Contig N50 (bp)	332,081	6,631	230,490
Contig number	31	5,653	11
BUSCO completeness (%)	85.7	88.5	99.5
Total length of repeats (Mb and %)	3.52 36.79	4.72 26.02	0.4 15.18
Number of protein-coding genes	2,847	4,231	2,074
Mean len. of protein- coding genes (bp)	1,049	952	1,019
Sequencing technology	ONT plus Illumina	Illumina	ONT plus Illumina
Publication	new	new	Mascarenhas dos Santos et al. (2023)

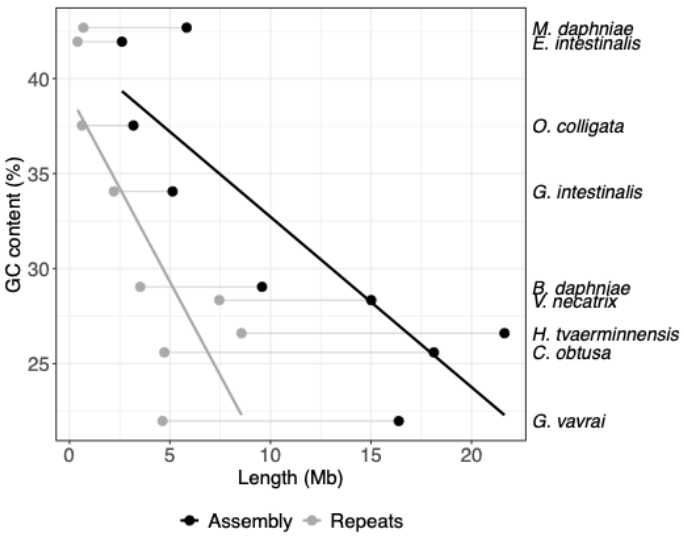

Figure 1: Relationship between genome length, total repeat length, and GC-content in nine microsporidia species. Assembly length (black dots and line) and total length of repeats (gray dots and line) correlated negatively with GC-content. Species names are displayed to the right of the plot at the corresponding GC-content.

Our improved *O. colligata* and *G. intestinalis* genome assemblies were more complete than previous assemblies, containing subtelomeric sequences for most chromosomes (Table S2). Phylogenetic inference confirmed previous rRNA gene phylogenies, placing our eight core microsporidia genomes into four different clades; *M. daphniae* was placed with the expanded microsporidium *Paramicrosporidium saccamoebae* (Figure 2 and S2). Our first high-quality genome assemblies in the Amblyosporida clade were all from the Gurleyidae family.

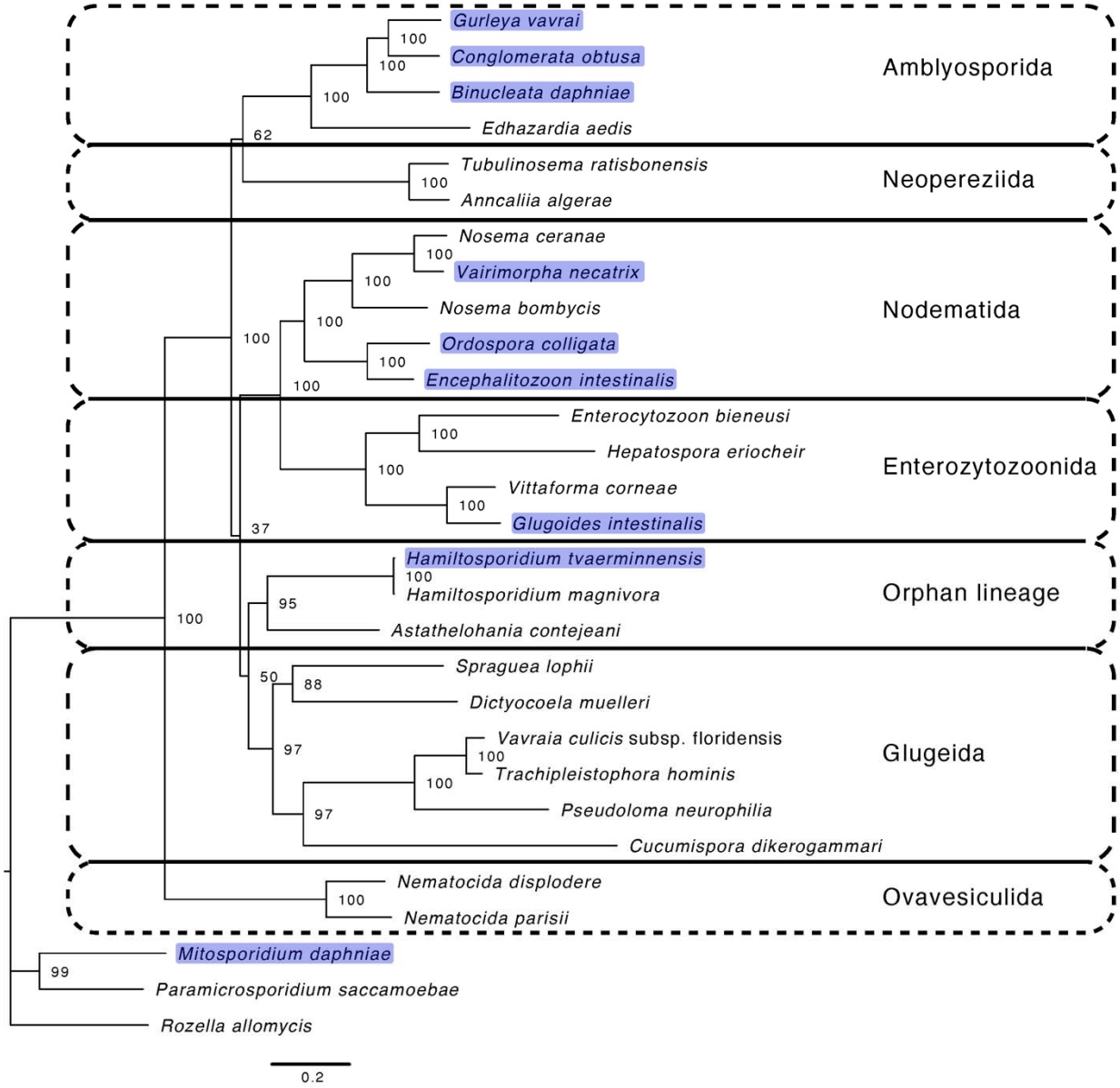

Figure 2: Maximum likelihood phylogeny of the core and expanded microsporidia with *Rozella allomycis* as an outgroup. In addition to the species discussed here (highlighted), we selected a representative sample of microsporidia species with available data for phylogenetic inference based on 15 single-copy orthologs. Node labels represent bootstrap values from 1,000 runs, and major microsporidia clades are delineated according to Bojko et al. (2022).

Genome architecture and synteny

Our reference and starting point for this comparison of microsporidia genome architecture was the well-described chromosome arrangement of *Encephalitozoon* (Mascarenhas dos Santos et al., 2023). We first compared *E. intestinalis* to its close relative *O. colligata*, finding them to have a similar genome architecture with a 5-mer

telomeric repeat (TTAGG) followed by a large and a small rRNA gene each (= rRNA operon), which we found on both ends of six out of the nine contigs and on one end of each of the other three contigs (Table S2). The next most closely related species, *V. necatrix* and *G. intestinalis*, had the same 5-mer telomeric repeat and a 4-mer telomeric repeat motif (TTAG), respectively. In *G. intestinalis*, rRNA genes followed the telomeric repeats on both ends of four contigs, on one end of seven contigs, and were absent in two contigs with telomeric repeats, indicating variation in chromosomal arrangement. The *V. necatrix* genome was more repetitive than its closest relatives, with more repeats between telomeres and rRNA genes, and with variation in their presence and number. In the other major clades containing Amblyosporida and *H. tvaerminnensis*, the genome architecture differed even more widely. The rRNA operons were not bound to the subtelomeres and occurred in tandem with tail-to-tail orientation. *Hamiltosporidium tvaerminnensis* had ten rRNA operons at five locations, and *G. vavrai* had 14 copies at seven locations. Each pair of juxtaposed rRNA operons surrounded specific LINE repeats that were predominantly found there. *G. vavrai* was also found to contain the previously described telomeric repeat motif of *H. tvaerminnensis* (TTAGGG; Angst et al. (2023)). From the less contiguous genome assemblies of *B. daphniae* and *C. obtusa*, we could recover no telomeric repeats and only a few to no rRNA genes (Table S2).

Overall, we found limited concordance in sequence arrangement (or synteny) across the analyzed microsporidia unless they were closely related (Figure 3). For example, although single-copy BUSCO genes, which are preserved due to their importance for organismal function and survival, occurred in conserved linkages across closely related species, only a few BUSCO linkage groups were shared by all core microsporidia. This was as expected due to the high degree of species divergence. Throughout the phylogeny, we found no centromeric regions based on chromosome-wide GC or repeat content, supporting hypotheses that microsporidian chromosomes might be holocentric, or their centromeres might be epigenetically defined (Khalaf et al., 2024; Mascarenhas dos Santos et al., 2023).

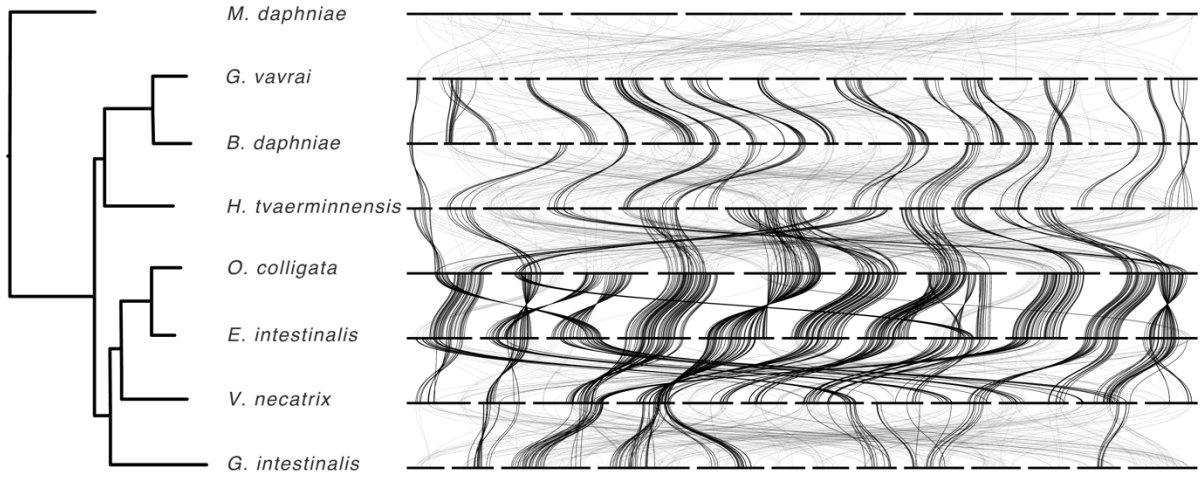

Figure 3: Ribbon diagram with maximum likelihood phylogeny showing synteny limited to closely related microsporidia. Contigs, some of which represent complete chromosomes, are represented by horizontal bars and connected by ribbons that connect the location of individual single-copy BUSCO genes across assemblies. Bold-type ribbons represent statistically conserved linkages. Because of its lower assembly contiguity, *Conglomerata obtusa* was omitted from the analysis.

Repeat landscapes

Encephalitozoon spp. have few to no repeats except at chromosome ends, which are all highly similar (Mascarenhas dos Santos et al., 2023). However, as seen in Figure 1, the genomes analyzed here showed high variation in repeat content, suggesting that repeats may play a role in shaping the evolution of microsporidian genomes. Using the Kimura substitution level, which estimates divergence from a repeat consensus sequence, we assessed repeat content and its activity over evolutionary time. High substitution levels indicate high divergence related to activity deep in the evolutionary past, and low substitution levels indicate low divergence and therefore recent activity. The largest genomes showed high repeat content and constant repeat activities throughout their past (Figure 4). The three Amblyosporida species shared similar repeat activity, especially in their early past, indicative of the activity in a common ancestor. DNA and LTR transposons were highly active in all three species, but *G. vavrai*, and especially *B. daphniae*, showed higher activity in the recent past. The apparent lack of recent activity in *C. obtusa* might be an artefact of the low divergence repeats collapsed in this short-read assembly. Recent increases in repeats compared to earlier activity were more pronounced in species with smaller genomes, such as LINEs in *G. intestinalis.* These observations all indicate that different transposon families could shape the larger genomes. As species with shorter genomes had less activity in their more distant past, repeats may have only recently been expanded or removed in the shorter genomes.

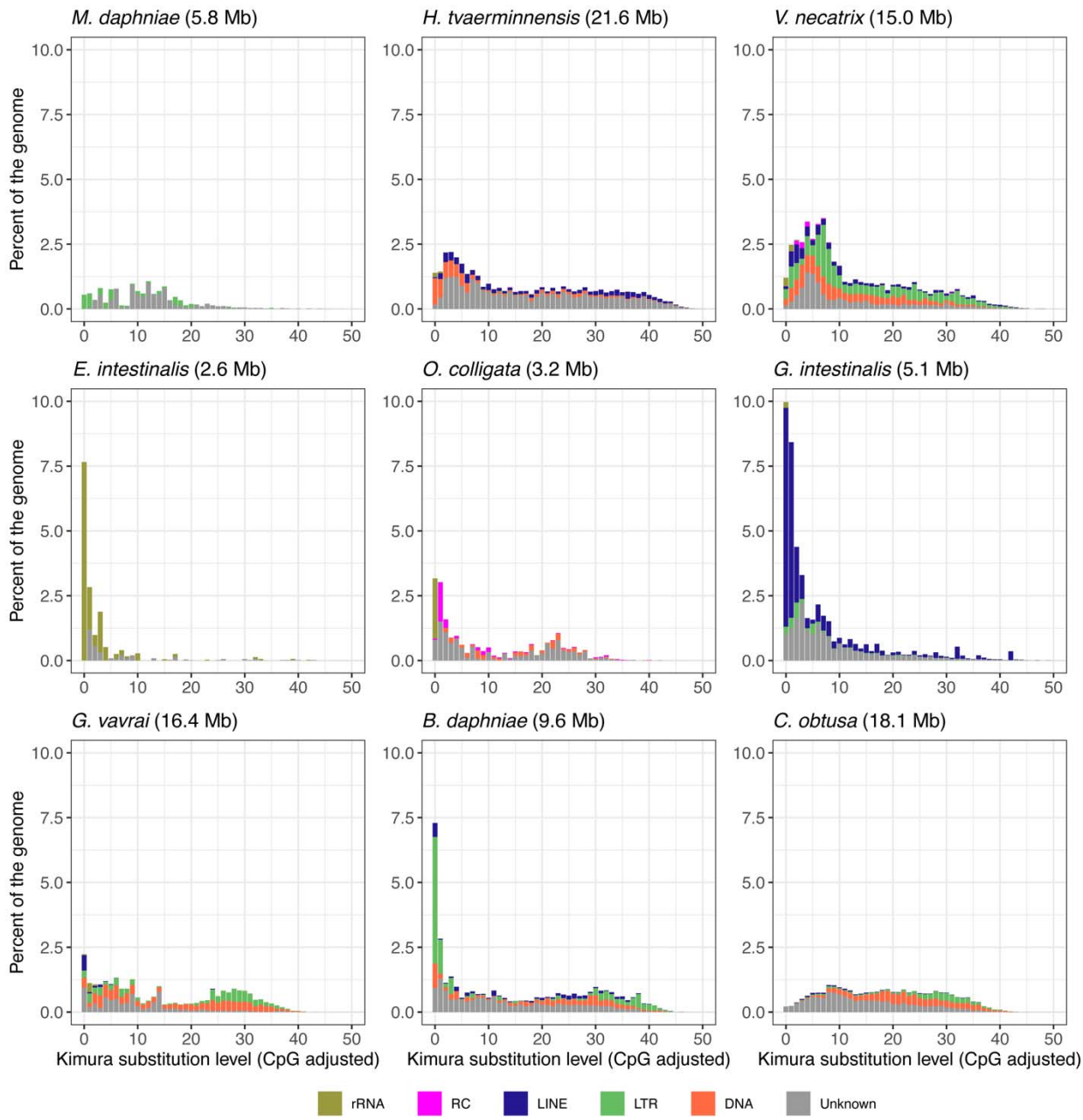

Figure 4: Repeat landscapes of microsporidia. Low Kimura substitution levels represent low divergence from the repeat consensus sequence, corresponding to recently expanded sequences, whereas high substitution levels indicate high divergence, indicating earlier activity. Genome assembly lengths are indicated in brackets.

Methylation

Epigenetic modifications may act to regulate not only gene expression but also repeat element activity and the associated variation in genome architecture. *Encephalitozoon*

has hypermethylated chromosome ends with rRNA genes and hypomethylated chromosome cores with mRNA genes (Mascarenhas dos Santos et al., 2023). Using a genome ontology approach based on hypergeometric tests, we compared these methylation patterns with the genomes for which we have epigenetic information. From our species with available ONT data for methylation analysis, we found that the closest *Encephalitozoon* relative, *G. intestinalis*, showed the most similar pattern: hypermethylated rRNA genes at the chromosome ends and hypomethylated mRNA genes at the chromosome cores (Tables 2 and S3). Unlike *Encephalitozoon*, however, *G. intestinalis* had LTRs and hypermethylated LINE repeats. In the other species with available ONT data, *H. tvaerminnensis*, *G. vavrai*, and *B. daphniae* also had hypermethylated rRNA genes and tended to have hypermethylated transposable repeat elements like DNA, LINE, and LTR, even though their genome architectures were very different than *Encephalitozoon* and *G. intestinalis*. Non-transposable repeats, like simple repeats and low-complexity regions, were hypomethylated in all species. Coding sequences were hypomethylated in the *H. tvaerminnensis* genome, like in *Encephalitozoon* and *G. intestinalis*, but were hypermethylated in *G. vavrai* and *B. daphniae*. Differential methylation of coding sequences and repeat elements could indicate gene expression and repeat activity regulation, respectively. The two repeat elements that expanded most recently (LINEs in *G. intestinalis* and LTRs in *B. daphniae*) were not consistently hyper- or hypomethylated.

Table 2: Methylation patterns of microsporidia. Overrepresented (= hypermethylated) genomic features have a negative log *P*-value, whereas underrepresented (= hypomethylated) genomic features have a positive log *P*-value. Asterisks denote levels of statistical significance obtained from genome ontology, while shading highlights directionality. Specifically, * indicates $p < 0.05$, ** indicates $p < 0.01$, and *** indicates $p < 0.001$.

	E. intestinalis	*B. daphniae*	*G. vavrai*	*G. intestinalis*	*H. tvaerminnensis*
DNA repeat	–	4.41 *	-21.86 ***	–	-202.4 ***
LINE repeat	–	5.06 **	1.46	-136.83 ***	-11.48 ***
Low complexity	2.90	4.44 *	21.08 ***	1	13.75 ***
LTR	–	-1.91	-91.96 ***	-1.58	-7.53 ***
mRNA	1,476.72 ***	-5.39 **	-4.91 **	77.31 ***	19.55 ***
RC repeat	–	–	–	–	-1.25
rRNA	-2,936.43 ***	-72.86 ***	-35.97 ***	-107.06 ***	-1.05
Satellite repeat	–	–	–	–	-2.73
Simple repeat	17.22 ***	14.72 ***	62.19 ***	7.99 ***	31.30 ***
Unknown repeat	-163.04 ***	-17.66 ***	-1.9	-200.41 ***	-8 ***

Present and absent protein domains

Genome streamlining in microsporidia has resulted in the loss of protein domains that are considered essential in non-parasitic free-living fungi (Haag et al., 2014; Jespersen et al., 2022). Using comparative genomics, we identified present and absent metabolic and functional capacities across microsporidia, looking at which functions were retained in all or most species and which were reduced. We began by summarizing the presence and absence of protein domains (Pfam)–characteristic protein families for biological pathways (Table S4). Among the nine microsporidia genomes, the extended microsporidium *M. daphniae* had, as expected, the most exclusive protein domains (656), even more than the shared protein domains across all nine species (444). These included mitochondrial carrier proteins (PF00153) that are characteristic of the transfer of molecules across mitochondrial membranes (*M. daphniae* has a mitochondrion) but that were absent in all mitochondria-free canonical microsporidia (Figure 5). Asparyl protease domains, e.g., PF13650, were among the most abundant protein domains with presence/absence polymorphism among the nine microsporidia (Figure 5). These endopeptidases are known virulence factors of pathogenic fungi (Mandujano-González et al., 2016). Looking at all species used in our phylogeny (Figure 2), we found that the functional glycolysis characteristic Fructose-bisphosphate aldolase class-I family (PF00274) was absent in *Enterocytozoon bieneusi* (Table S4), consistent with previous findings that glycolysis is reduced in this species that obtains its energy directly from the host (Wiredu Boakye et al., 2017). Protein domains like ABC transporter (PF00005), Hsp70 (PF00012), and actin (PF00022) were present in all species (Table S4) and thus appear to be essential for microsporidia.

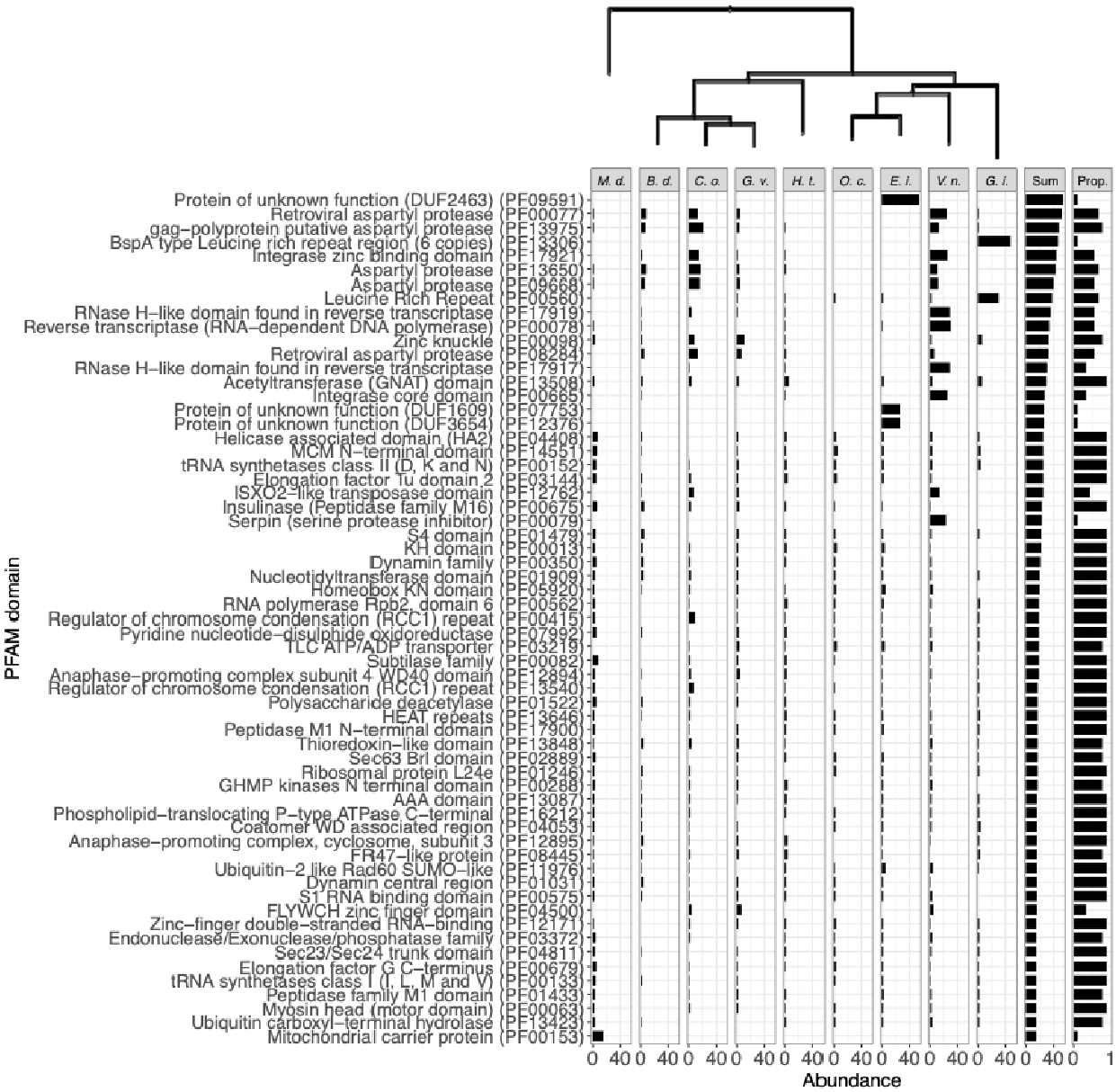

Figure 5: Variably present protein domains (Pfam) ordered from most to least abundant. Rows list individual protein domains with their Pfam domain description and ID (in parentheses), while columns with a maximum likelihood phylogeny at the top list the nine species. The penultimate column tallies the number of protein domains present, while the last column provides the proportion of species in which they occur. Full data is presented in Table S4.

Discussion

Microsporidia are intracellular parasites that are enigmatic due to their high degree of specialization. Our goal was to better understand the evolution of their genome by producing several new high-quality genome assemblies that reveal the (epi)genetic architecture across this taxon. We focused mainly on defining which features are general for all microsporidia and which are derived in subclades. GC-content, genome length, chromosomal sequence arrangement, and methylation patterns vary widely among microsporidian clades; however, species that share a more recent common ancestor tend to exhibit less variation in these features, similar to other systems with high-quality genomes available for study. The architecture of microsporidia genomes studied here were only similar to the previously described *Encephalitozoon* in one close relative and in a species of the next closest major microsporidian clade, proving that *Encephalitozoon*'s chromosomal sequence arrangement is not general to all microsporidia, as has been speculated (Brugère et al., 2000). Our other studied species showed very diverse chromosomal arrangements and repeat contents, although the same genetic features tend to be hyper- or hypomethylated in all species. For example, rRNA gene positions are different among all microsporidian clades, but always hypermethylated in all species. Because the new and improved microsporidia genome assemblies presented here are all from species that infect the aquatic crustacean *Daphnia,* we can see that the variation in genome architecture is not explained by the host: indeed, the host's role in parasite genome architecture appears to be minor. As models, thus, *Daphnia* and its microsporidian parasites from across the phylum allow us to extend microsporidia research and improve our understanding of genome evolution in these extreme parasites beyond the mammal-parasite model.

Nucleotide composition linked to genome size

Nucleotide composition is fundamental to genome architecture, and although GC-content varies widely across all lifeforms, it tends to evolve to lower values in parasites such as the microsporidia. Reduced recombination rates, diminished DNA repair mechanisms, relaxed purifying selection, and increased selection for AT mutations have all been proposed as reasons for reduced GC in parasite genomes (Videvall, 2018). These mechanisms may play a role in the negative correlation observed here between GC-content and repeat content, which is also seen in fungi and other related phyla (Elliott & Gregory, 2015). First, it might be easier for repeat elements to proliferate in species with relaxed purifying selection due to the small effective population size relative to the mode of transmission (Haag et al., 2020). We expect species with a small effective population size to have more and more recent expansions of repeat elements due to vertical transmission compared to species with a higher effective population size and horizontal-only transmission (de Albuquerque et al., 2020). With our sample, it is hard to draw conclusions on demographic effect; however, long genome Amblyosporida and Nosema/Vairimorpha clade species are known for complex life cycles with mixed-mode transmission (Vávra et al., 2018; Xiong et al., 2023), and mixed-mode transmission in the large genome microsporidium *H. tvaerminnensis* has been associated with small effective population sizes (Haag et al., 2020). Contrarily, species with small genomes are only transmitted horizontally. Second, AT mutations might accumulate in proliferating repeat elements, directly and indirectly accelerating the

decrease in GC-content. This occurs indirectly if 5mC methylation is used to slow the proliferation of repeat elements because fewer positions are available for this epigenetic modification. In contrast, GC-richer repeats might be more of a methylation target, possibly constraining their proliferation. Although repeat elements were hypermethylated across species, repeat activity was more evenly distributed across time in lower GC species than in higher GC species, where activity occurred mainly in the recent past. This could be because higher GC species have more stringent selection against repeat proliferation, more efficient mechanisms to remove repeats, higher recombination rates, or lower tendency for AT mutation, such that our methodology estimates higher recent repeat activity because of the lower divergence among them.

Distribution of rRNA genes

The rRNA genes are of special interest in microsporidia genome architecture, mainly because of their unusual location in *Encephalitozoon* close to each chromosome end, at the subtelomeres (Mascarenhas dos Santos et al., 2023). Compared to previous findings from a short-read-based draft genome (Mascarenhas dos Santos et al., 2023; Pombert et al., 2015), we show that *O. colligata* shares this structure with its close *Encephalitozoon* relatives. This highlights the use of high-quality resources to study genome architecture. rRNA genes are also located at some, but not all, of the subtelomeres in a species of the next closely related major clade, *G. intestinalis,* which also shows small-scale rRNA gene rearrangements (Refardt & Mouton, 2007). Given that transposable elements (TEs) show high and recent activity in *G. intestinalis*, that rRNA genes and TEs are co-located, and that TEs interact with and benefit from each other (Garcia et al., 2024), we suggest that rRNA genes and TEs may coevolve together in this microsporidium or even across microsporidia, a hypothesis supported also by our finding that rRNA genes and TEs are co-located in *H. tvaerminnensis* and Amblyosporida microsporidia. Moreover, rRNA genes in these species have yet another outstanding arrangement: rRNA operons occur in tandem with tail-to-tail orientation and few LINE or unknown repeats between them. These rRNA operon pairs are not located close to a chromosome end, however, suggesting that the proximity of rRNA genes to chromosome ends is a derived characteristic found in some species of the Nosematida and Enterozytozoonida clades.

Co-occurrence and epigenetic modification of repeat elements

The significance of methylation is unclear in microsporidia. Mascarenhas dos Santos et al. (2023) have speculated that methylation at rRNA genes facilitates heterochromatin formation, helping to silence the ribosomal machinery during phases of the life cycle that have low or no access to energy. Indeed, we found rRNA genes to be hypermethylated in all the species studied, regardless of their location. The hypothesis proposed by Mascarenhas dos Santos et al. (2023) could also apply to other genetic loci, such as TEs, which because their activity may be unfavorable, are silenced by epigenetic modifications that facilitate heterochromatin formation. Whereas small microsporidia genomes have low to no repeats, including TEs, and hypomethylated

cores, larger genomes might show methylation at repeat loci across the genome to reduce TE activity through heterochromatin formation. Given that rRNA genes and transposable elements co-occur in the studied microsporidia, hypermethylation could help to simultaneously control their activity and spread. At the same time, their co-occurrence might be beneficial to the TEs, which might be more likely to proliferate or less likely to be removed if they occur close to the rRNA genes, which are highly expressed. The microsporidia *V. necatrix* and *Nosema muscidifuracis* from the same clade as *Encephalitozoon* and *Ordospora* have more randomly distributed rRNA operons (and *N. muscidifuracis* has many more), while both have much lower whole-genome GC-content than their relatives (Xiong et al., 2023). A telomere-to-telomere assembly of the *N. muscidifuracis* genome would help pin down rRNA gene location and methylation data for the two species to clarify the role of epigenetic modification in the evolution of microsporidia repeat families. Neither epigenetic, nucleotide, nor sequence composition has indicated the location of centromeres in the species studied here. Hi-C sequencing could help identify centromeres (Khalaf et al., 2024).

Conclusion

Microsporidia have attracted much attention because of their parasitic life history and because they include some of the smallest known eukaryote genomes. However, their genomes also exhibit drastic variation in size and other features. Here we present high-quality genome assemblies from across the major microsporidia clades, showing covariance in GC-content, chromosomal sequence arrangement, and epigenetic modification. Species with high GC-content have a sequence arrangement similar to *Encephailtozoon* spp., with hypermethylated subtelomeres and hypomethylated cores. In contrast, low-GC species do not share this pattern, having instead a higher activity of repeat elements, which have lower GC-content than coding sequences, likely because the latter are under purifying selection (Videvall, 2018). Low-GC genomes might arise due to a lowered selection regime that allow transposable repeats to proliferate more easily (de Albuquerque et al., 2020; Videvall, 2018). The evolutionary dynamics leading to the discrepancy between relatively strong selection on coding sequences and weak selection outside of them remain unknown. Still, the activity of repeat elements clearly shapes the evolution of genome architecture, and with the recent advent of high-throughput sequencing technologies should thus be included in comparative studies of GC-content where it was previously not.

Although the streamlining of genomes has led to a significantly fewer coding sequences across all core microsporidia (Žárský et al., 2023), certain species show extraordinary genome reductions, like the partial absence of glycolysis genes in *Enterocytozoon* spp. (Wiredu Boakye et al., 2017). Here, we provide an overview of the presence and absence of protein domains across microsporidia as a resource for studies focused on the presence/absence polymorphism and functional adaptations to parasitism across the microsporidia clades. Knowing the phylogenetic distribution of protein domains and other genomic features in the microsporidia allows us to dive deeper into the evolution of this extraordinary group of parasites. Features that are conserved by all microsporidia may have contributed to the radiation of this group, with the loss of typical mitochondria

being the hallmark example. In contrast, features present only in individual microsporidian clades or species may have been important in evolving adaptations to specific hosts or life histories. Furthermore, correlations between the presence of certain genomic features and life history traits can shed light on the evolutionary drivers of these traits. Thus, the presence or absence of protein domains and other genomic features can be the starting point for in-depth comparative study of the microsporidia and their genomic architecture with these high-quality genomes. As a final step for obtaining telomere-to-telomere assemblies, especially the harder-to-assemble longer genomes, we suggest applying Hi-C sequencing to microsporidia.

Methods

Samples

Using long wide-mouth pipettes, we collected *Daphnia pulex* infected with *C. obtusa* from the Tvaerminne archipelago, Finland (59°49'55.4", N 23°15'17.6" E), placed them in RNAlater (Ambion, Glasgow, United Kingdom), and kept them refrigerated until DNA isolation. *Daphnia pulex* infected with *G. vavrai* were collected from Aegelsee lake near Frauenfeld, Switzerland (47°33′28.0″ N, 8°51′46.0″ E) and were transported to the laboratory for DNA isolation. *Daphnia magna* infected with *B. daphniae* (ID: BE-OM-3), *G. intestinalis* (KZ-23-1), and *O. colligata* (GB-LK1-1) were previously collected and kept in the lab as iso-female lines (Angst et al., in prep.; Pombert et al., 2015; Refardt et al., 2008) until DNA isolation. We obtained high-molecular-weight DNA from *Daphnia* with parasites using the GenePure DNA Isolation Kit (QIAGEN, Hilden, Germany) as described in Angst et al. (in review) (https://dx.doi.org/10.17504/protocols.io.5jyl82n96l2w/v1; Angst & Fields (2024)). Illumina paired-end sequencing and either PacBio HiFi (Pacific Biosciences high fidelity) or ONT (Oxford Nanopore Technologies) sequencing was used to obtain the highest quality nucleotide or modification calls, respectively. Illumina libraries were prepared using Kapa PCR-free kits and sequenced by the Quantitative Genomics Facility service platform at the Department of Biosystem Science and Engineering (D-BSSE, ETH) Basel, Switzerland, on an Illumina HiSeq 6000 sequencer. PacBio SMRTbell libraries were prepared and sequenced on a PacBio Revio at the Lausanne Genomics Technologies Facility (GTF, UNIL), Switzerland. We prepared ONT libraries with the SQK-LSK110 ligation kit and sequenced them using a MinION device with a Spot-ON Flow Cell (R9.4.1). Additionally, we downloaded genomic data of *E. intestinalis* (NCBI database; Assembly name: ASM2439929v1; GenBank assembly accession: GCA_024399295.1; SRA accession: SRR17865590; Bioproject accession: PRJNA594722; Mascarenhas dos Santos et al. 2023) *H. tvaerminnensis* (NCBI database; Assembly name: FIOER33 v3; GenBank assembly accession: GCA_022605425.2; SRA accession: SRR24575619, Bioproject accession: PRJNA778105; Angst et al. 2023), *M. daphniae* (NCBI database; Bioproject accession: PRJNA1199805; Angst et al. (in prep.)), and *V. necatrix* (NCBI database; Assembly

name: ASM3663032v1; GenBank assembly accession: GCF_036630325.1; Bioproject accession: PRJNA909071; Svedberg et al. 2024) for analysis.

Assembly

Glugoides intestinalis and *O. colligata*

Glugoides intestinalis (total/average read length: 79.38 Gb/10.75 Kb) and *O. colligata* (20.22 Gb/8.84 Kb) were sequenced using PacBio HiFi sequencing to an average sequencing coverage of 55× and 45×, respectively. We used hifiasm v.0.19.8-r603 (Cheng et al., 2021) to assemble the obtained sequencing reads with default parameters for *G. intestinalis* and with --hom-cov 45 -l1 --n-hap 1 parameters for *O. colligata*. Additionally, we generated a second hifiasm assembly for both species after excluding host-specific PacBio HiFi reads using SAMtools v.1.7 (Danecek et al., 2021) that were identified by mapping them to the *D. magna* genome (NCBI database; Assembly name: ASM4014379v1; GenBank assembly accession: GCA_040143795.1; BioProject ID: PRJNA624896; Cornetti et al. 2024) using minimap2 v.2.20-r1061 (H. Li, 2018). For each species, we aligned the two assemblies using minimap2 and merged them manually. We assessed the quality and completeness of the assemblies using QUAST v.5.0.2 (Gurevich et al., 2013) and BUSCO v.5.5.0 (Manni et al., 2021) with the microsporidia_odb10 database (Creation date: 2020-08-05). All software was used with default parameters unless stated otherwise.

Binucleata daphniae and *G. vavrai*

We sequenced *B. daphniae* (total/N50: 0.72 Gb/7.21 Kb) and *G. vavrai* (3.96 Gb/3.53 Kb) using ONT sequencing in adaptive mode, depleting for sequences homologous to the *D. magna* and the *D. pulex* genome (NCBI database; Assembly name: PA42 4.2; GenBank assembly accession: GCA_911175335.1; BioProject ID: PRJEB46221; Ye et al. 2021), respectively. We used bonito v.0.5.3 (https://github.com/nanoporetech/bonito) to base-call ONT sequencing reads with the dna_r9.4.1_e8.1_sup@v3.3 model. For the *B. daphniae* genome assembly, we used MaSuRCA v.4.0.5 (Zimin et al., 2017) with ONT (53×) and Illumina (639×) sequencing reads, excluding host-specific, adaptor, and low-quality sequences using minimap2, bwa-mem2 mem v.2.2.1 (Vasimuddin et al., 2019), SAMtools, and fastp v.0.23.2 (Chen, 2023) with -q 30 and -l 100 parameters. Haplotigs were purged twice from the assembly using the Purge Haplotigs pipeline v.1.1.2 (Roach et al., 2018) with -l 125 -m 220 -h 400 parameters and BEDTools v.2.30.0 (Quinlan, 2014) for generating the read-depth distribution. For the *G. vavrai* genome assembly, we used nextDenovo v.2.5.0 (Hu et al., 2024) with ONT sequencing reads (98×) and polished the resulting assembly with Illumina sequencing reads (458×) using nextPolish v.1.4.1 (Hu et al., 2020), Medaka v.1.7.2 (https://github.com/nanoporetech/medaka) and Pilon v.1.24 (Walker et al., 2014). The Purge Haplotigs pipeline was applied with -l 88 -m 176 -h 300 parameters.

Conglomerata obtusa

Because we were unable to generate sufficient quality genomic material to conduct long-read sequencing for *C. obtusa*, we used Illumina sequencing (153×) for this case only. We excluded sequencing reads mapping to the *D. pulex* host genome and assembled the remaining reads using megahit v.1.2.9 (D. Li et al., 2016). We excluded contigs with less than a third of the average coverage, shorter than 500 basepairs, with GC-content above 33, or with coverage over 1× for a *Daphnia* sample from a parasite-free pond of the same island (NCBI database; SRA accession: SRR26394761, Bioproject ID: PRJNA862292; Angst et al. 2024) using seqtk v.1.3 (https://github.com/lh3/seqtk) and BBMap v.38.96 (Bushnell, 2014).

Genome annotation and phylogenetic inference

We annotated *de novo* genome assemblies as described in Angst et al. (in prep.) using funannotate v.1.8.14 (Palmer & Stajich, 2022) and associated software for gene prediction and functional annotation (Almagro Armenteros et al., 2019; Brůna et al., 2020; Cantalapiedra et al., 2021; Frith, 2011; Haas et al., 2008; Huerta-Cepas et al., 2019; Jones et al., 2014; Käll et al., 2004; Korf, 2004; Majoros et al., 2004; Seppey et al., 2019; Stanke et al., 2006, 2008). For phylogenetic inference, we downloaded genomes from a broad range of microsporidia (Aurrecoechea et al., 2017) including the extended microsporidium *P. saccamoebae* (Quandt et al., 2017) along with *Rozella allomycis* (James et al., 2013) as an outgroup. We identified single-copy orthologs among the species using proteinortho v.6.3.0 (Klemm et al., 2023), concatenated them, aligned the individual genes as well as the concatenated genes using MAFFT v.7.508 (Katoh et al., 2002; Katoh & Standley, 2013), and trimmed the alignments using trimAl v.1.4.rev15 (Capella-Gutiérrez et al., 2009) with the -automated1 flag. We inferred a maximum likelihood species tree from the concatenated sequences using IQ-TREE 2 v.2.2.0 (Minh et al., 2020) and a weighted phylogeny from the individual maximum likelihood gene trees using Weighted ASTRAL v.1.19.3.7 (Zhang & Mirarab, 2022) with -r 16 -s 16 -x 100 -n 0 parameters separately for all species and the nine main study species.

Comparative genomics and genome ontology

We looked for homology and synteny across genomes using odp v.0.3.3 (Schultz et al., 2023). Specifically, we used the coordinates of the identified BUSCO genes to infer ancestral BUSCO gene linkage groups, test for conserved linkages among genomes, and create the ribbon diagram. To obtain modified base calls from ONT sequencing reads, we used megalodon v.2.5.0 (https://nanoporetech.github.io/megalodon) focusing on 5mC modifications for downstream analyses. We tested whether genomic annotations, including mRNA genes, rRNA genes, and different repeat types, were enriched with methylation sites by using the genome ontology peak annotation enrichment in homer v.5.1 (Heinz et al., 2010). This analysis considered sites to be methylated if read depth was above 3× and percent modified reads over 50 %. The inclusion of multiple cell stages in our DNA isolation protocol explains, at least in part, variation in epigenetic modification between sequencing reads. For example, spores

might have a different methylation pattern than sporonts, meronts, or other life stages (Mascarenhas dos Santos et al., 2023). For *G. intestinalis*, we re-used ONT reads from Angst et al. (in prep.) (NCBI database; SRA accession: SRR31799445, Bioproject accession: PRJNA1199805). We identified repeats using RepeatModeler v.2.0.2, including the LTR pipeline (Flynn et al., 2020), and ReapeatMasker v.4.1.2 (Smit et al., 2013), and separately identified rRNA genes using cmscan of the Infernal pipeline v.1.1.2 (Nawrocki & Eddy, 2013). From the ReapeatMasker alignments, we created repeat landscapes using ReapeatMasker utils calcDivergenceFromAlign.pl and createRepeatLandscape.pl. Finally, we summarized PFAM domains to identify the presence and absence of protein domains across all species using the module compare of funannotate, and the tidyverse 2.0.0 (Wickham et al., 2019) R v.4.4.1 (R Core Team, 2024) package for representation and statistics.

Data availability

Raw data are deposited in the NCBI SRA database, and the assembled genome plus the predicted protein sequences are available in the NCBI GenBank (BioProject ID: PRJNA1199946).

Acknowledgments

We thank Ju□rgen Hottinger, Michelle Krebs, and Urs Stiefel for their help in the laboratory and all members of the Ebert group for providing feedback on the study and the manuscript. Alix Thivolle extracted DNA from *D. pulex* infected with *C. obtusa*. Christina Tadiri collected *D. pulex* infected with *G. vavrai*. Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing center at the University of Basel. Suzanne Zweizig improved the language of the text.

Funding

This work was supported by the Swiss National Science Foundation (SNSF) (grant numbers 310030_188887 and 310030_219529 to D.E.).

Conflicts of interest

None declared.

References

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

Angst, P., & Fields, P. D. (2024). DNA-extraction of Daphnia and symbionts. *Protocols.Io*. https://doi.org/10.17504/protocols.io.5jyl82n96l2w/v1

Angst, P., Haag, C. R., Ben-Ami, F., Fields, P. D., & Ebert, D. (2024). Genome-wide allele frequency changes reveal that dynamic metapopulations evolve differently. *Molecular Biology and Evolution*, *41*(7), msae128. https://doi.org/10.1093/molbev/msae128

Angst, P., Pombert, J.-F., Ebert, D., & Fields, P. D. (2023). Near chromosome–level genome assembly of the microsporidium *Hamiltosporidium tvaerminnensis*. *G3 Genes|Genomes|Genetics*, jkad185. https://doi.org/10.1093/g3journal/jkad185

Angst, P., Thivolle, A., Haden, Z., Wale, N., & Ebert, D. (in review). *Genomic analysis of the zooplankton-associated pathogenic bacterium Spirobacillus cienkowskii reveals its functional and metabolic capacities*.

Angst, P., Thivolle, A., Krebs, M., Galimov, Y., Fields, P. D., & Ebert, D. (in prep.). *Phenotypic and molecular convergent evolution of microsporidia infecting host gut epithelium*.

Aurrecoechea, C., Barreto, A., Basenko, E. Y., Brestelli, J., Brunk, B. P., Cade, S., Crouch, K., Doherty, R., Falke, D., Fischer, S., Gajria, B., Harb, O. S., Heiges, M., Hertz-Fowler, C., Hu, S., Iodice, J., Kissinger, J. C., Lawrence, C., Li, W., … Zheng, J. (2017). EuPathDB: The eukaryotic pathogen genomics database resource. *Nucleic Acids Research*, *45*(Database issue), D581–D591. https://doi.org/10.1093/nar/gkw1105

Bojko, J., Reinke, A. W., Stentiford, G. D., Williams, B., Rogers, M. S. J., & Bass, D. (2022). Microsporidia: A new taxonomic, evolutionary, and ecological synthesis. *Trends in Parasitology*, *38*(8), 642–659. https://doi.org/10.1016/j.pt.2022.05.007

Bojko, J., & Stentiford, G. D. (2022). Microsporidian pathogens of aquatic animals. In L. M. Weiss & A. W. Reinke (Eds.), *Microsporidia: Current Advances in Biology* (pp. 247–283). Springer International Publishing. https://doi.org/10.1007/978-3-030-93306-7_10

Brugère, J.-F., Cornillot, E., Méténier, G., Bensimon, A., & Vivarès, C. P. (2000). *Encephalitozoon cuniculi* (Microspora) genome: Physical map and evidence for telomere-associated rDNA units on all chromosomes. *Nucleic Acids Research*, *28*(10), 2026–2033.

Brůna, T., Lomsadze, A., & Borodovsky, M. (2020). GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics*, *2*(2), lqaa026. https://doi.org/10.1093/nargab/lqaa026

Bushnell, B. (2014). *BBMap: A Fast, Accurate, Splice-Aware Aligner* (No. LBNL-7065E). Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). https://www.osti.gov/biblio/1241166-bbmap-fast-accurate-splice-aware-aligner

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. https://doi.org/10.1093/molbev/msab293

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta*, *2*(2), e107. https://doi.org/10.1002/imt2.107

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), Article 2. https://doi.org/10.1038/s41592-020-01056-5

Cornetti, L., Fields, P. D., Du Pasquier, L., & Ebert, D. (2024). Long-term balancing selection for pathogen resistance maintains trans-species polymorphisms in a planktonic crustacean. *Nature Communications*, *15*(1), 5333. https://doi.org/10.1038/s41467-024-49726-8

Corradi, N., & Slamovits, C. H. (2011). The intriguing nature of microsporidian genomes. *Briefings in Functional Genomics*, *10*(3), 115–124. https://doi.org/10.1093/bfgp/elq032

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

de Albuquerque, N. R. M., Ebert, D., & Haag, K. L. (2020). Transposable element abundance correlates with mode of transmission in microsporidian parasites. *Mobile DNA*, *11*, 19. https://doi.org/10.1186/s13100-020-00218-8

Ebert, D. (2008). Host–parasite coevolution: Insights from the *Daphnia*–parasite model system. *Current Opinion in Microbiology*, *11*(3), 290–301. https://doi.org/10.1016/j.mib.2008.05.012

Elliott, T. A., & Gregory, T. R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1678), 20140331. https://doi.org/10.1098/rstb.2014.0331

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Frith, M. C. (2011). A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Research*, *39*(4), e23. https://doi.org/10.1093/nar/gkq1212

Garcia, S., Kovarik, A., Maiwald, S., Mann, L., Schmidt, N., Pascual-Díaz, J. P., Vitales, D., Weber, B., & Heitkam, T. (2024). The dynamic interplay between ribosomal DNA and transposable elements: A perspective from genomics and cytogenetics. *Molecular Biology and Evolution*, *41*(3), msae025. https://doi.org/10.1093/molbev/msae025

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Haag, K. L., James, T. Y., Pombert, J.-F., Larsson, R., Schaer, T. M. M., Refardt, D., & Ebert, D. (2014). Evolution of a morphological novelty occurred before genome compaction in a lineage of extreme parasites. *Proceedings of the National Academy of Sciences*, *111*(43), 15480–15485. https://doi.org/10.1073/pnas.1410442111

Haag, K. L., Pombert, J.-F., Sun, Y., de Albuquerque, N. R. M., Batliner, B., Fields, P., Lopes, T. F., & Ebert, D. (2020). Microsporidia with vertical transmission were likely shaped by nonadaptive processes. *Genome Biology and Evolution*, *12*(1), 3599–3614. https://doi.org/10.1093/gbe/evz270

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology*, *9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, *38*(4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Hu, J., Fan, J., Sun, Z., & Liu, S. (2020). NextPolish: A fast and efficient genome polishing tool for long-read assembly. *Bioinformatics*, *36*(7), 2253–2255. https://doi.org/10.1093/bioinformatics/btz891

Hu, J., Wang, Z., Sun, Z., Hu, B., Ayoola, A. O., Liang, F., Li, J., Sandoval, J. R., Cooper, D. N., Ye, K., Ruan, J., Xiao, C.-L., Wang, D., Wu, D.-D., & Wang, S. (2024). NextDenovo: An efficient error correction and accurate assembly tool for noisy long reads. *Genome Biology*, *25*(1), 107. https://doi.org/10.1186/s13059-024-03252-4

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

James, T. Y., Pelin, A., Bonen, L., Ahrendt, S., Sain, D., Corradi, N., & Stajich, J. E. (2013). Shared signatures of parasitism and phylogenomics unite Cryptomycota

and microsporidia. *Current Biology: CB*, *23*(16), 1548–1553. https://doi.org/10.1016/j.cub.2013.06.057

Jespersen, N., Monrroy, L., & Barandun, J. (2022). Impact of genome reduction in microsporidia. In L. M. Weiss & A. W. Reinke (Eds.), *Microsporidia: Current Advances in Biology* (pp. 1–42). Springer International Publishing. https://doi.org/10.1007/978-3-030-93306-7_1

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Käll, L., Krogh, A., & Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, *338*(5), 1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Khalaf, A., Francis, O., & Blaxter, M. L. (2024). Genome evolution in intracellular parasites: Microsporidia and Apicomplexa. *Journal of Eukaryotic Microbiology*, e13033. https://doi.org/10.1111/jeu.13033

Klemm, P., Stadler, P. F., & Lechner, M. (2023). Proteinortho6: Pseudo-reciprocal best alignment heuristic for graph-based detection of (co-)orthologs. *Frontiers in Bioinformatics*, *3*. https://doi.org/10.3389/fbinf.2023.1322477

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*(1), 59. https://doi.org/10.1186/1471-2105-5-59

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., & Lam, T.-W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, *102*, 3–11. https://doi.org/10.1016/j.ymeth.2016.02.020

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics*, *20*(16), 2878–2879. https://doi.org/10.1093/bioinformatics/bth315

Mandujano-González, V., Villa-Tanaca, L., Anducho-Reyes, M. A., & Mercado-Flores, Y. (2016). Secreted fungal aspartic proteases: A review. *Revista Iberoamericana de Micología*, *33*(2), 76–82. https://doi.org/10.1016/j.riam.2015.10.003

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Mascarenhas dos Santos, A. C., Julian, A. T., Liang, P., Juárez, O., & Pombert, J.-F. (2023). Telomere-to-Telomere genome assemblies of human-infecting *Encephalitozoon* species. *BMC Genomics*, *24*(1), 237. https://doi.org/10.1186/s12864-023-09331-3

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Nakjang, S., Williams, T. A., Heinz, E., Watson, A. K., Foster, P. G., Sendra, K. M., Heaps, S. E., Hirt, R. P., & Embley, T. M. (2013). Reduction and expansion in microsporidian genome evolution: New insights from comparative genomics. *Genome Biology and Evolution*, *5*(12), 2285–2303. https://doi.org/10.1093/gbe/evt184

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933–2935. https://doi.org/10.1093/bioinformatics/btt509

Palmer, J., & Stajich, J. (2022). *nextgenusfs/funannotate: Funannotate* (Version 1.8.11) [Computer software]. Zenodo. https://zenodo.org/record/2604804

Parisot, N., Pelin, A., Gasc, C., Polonais, V., Belkorchia, A., Panek, J., El Alaoui, H., Biron, D. G., Brasset, E., Vaury, C., Peyret, P., Corradi, N., Peyretaillade, É., & Lerat, E. (2014). Microsporidian genomes harbor a diverse array of transposable elements that demonstrate an ancestry of horizontal exchange with metazoans. *Genome Biology and Evolution*, *6*(9), 2289–2300. https://doi.org/10.1093/gbe/evu178

Pombert, J.-F., Haag, K. L., Beidas, S., Ebert, D., & Keeling, P. J. (2015). The *Ordospora colligata* Genome: Evolution of Extreme Reduction in Microsporidia and Host-To-Parasite Horizontal Gene Transfer. *mBio*, *6*(1), e02400-14. https://doi.org/10.1128/mBio.02400-14

Poulin, R., & Randhawa, H. S. (2015). Evolution of parasitism along convergent lines: From ecology to genomics. *Parasitology*, *142*(Suppl 1), S6–S15. https://doi.org/10.1017/S0031182013001674

Quandt, C. A., Beaudet, D., Corsaro, D., Walochnik, J., Michel, R., Corradi, N., & James, T. Y. (2017). The genome of an intranuclear parasite, *Paramicrosporidium saccamoebae*, reveals alternative adaptations to obligate intracellular parasitism. *eLife*, *6*, e29594. https://doi.org/10.7554/eLife.29594

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. (2024). *R: The R Project for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Refardt, D., Decaestecker, E., Johnson, P. T. J., & Vávra, J. (2008). Morphology, Molecular Phylogeny, and Ecology of *Binucleata daphniae* n. G., n. Sp. (Fungi: Microsporidia), a Parasite of *Daphnia magna* Straus, 1820 (Crustacea: Branchiopoda). *Journal of Eukaryotic Microbiology*, *55*(5), 393–408. https://doi.org/10.1111/j.1550-7408.2008.00341.x

Refardt, D., & Mouton, L. (2007). Reverse Arrangement of rRNA Subunits in the Microsporidium *Glugoides intestinalis*. *Journal of Eukaryotic Microbiology*, *54*(1), 83–85. https://doi.org/10.1111/j.1550-7408.2006.00149.x

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*(1), 460. https://doi.org/10.1186/s12859-018-2485-7

Schmid-Hempel, P. (2021). *Evolutionary Parasitology: The Integrated Study of Infections, Immunology, Ecology, and Genetics* (Second Edition, Second Edition). Oxford University Press.

Schultz, D. T., Haddock, S. H. D., Bredeson, J. V., Green, R. E., Simakov, O., & Rokhsar, D. S. (2023). Ancient gene linkages support ctenophores as sister to other animals. *Nature*, *618*(7963), 110–117. https://doi.org/10.1038/s41586-023-05936-6

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In M. Kollmar (Ed.), *Gene Prediction: Methods and Protocols* (pp. 227–245). Springer. https://doi.org/10.1007/978-1-4939-9173-0_14

Smit, A., Hubley, R., & Green, P. (2013). *RepeatMasker Open-4.0* [Computer software]. http://www.repeatmasker.org

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*(5), 637–644. https://doi.org/10.1093/bioinformatics/btn013

Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, *7*(1), 62. https://doi.org/10.1186/1471-2105-7-62

Svedberg, D., Winiger, R. R., Berg, A., Sharma, H., Tellgren-Roth, C., Debrunner-Vossbrinck, B. A., Vossbrinck, C. R., & Barandun, J. (2024). Functional annotation of a divergent genome using sequence and structure-based similarity. *BMC Genomics*, *25*(1), 6. https://doi.org/10.1186/s12864-023-09924-y

Vasimuddin, Md., Misra, S., Li, H., & Aluru, S. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *2019 IEEE International*

Parallel and Distributed Processing Symposium (IPDPS), 314–324. https://doi.org/10.1109/IPDPS.2019.00041

Vávra, J., Fiala, I., Krylová, P., Petrusek, A., & Hyliš, M. (2018). Molecular and structural assessment of microsporidia infecting daphnids: The “obtusa-like” microsporidia, a branch of the monophyletic Agglomeratidae clade, with the establishment of a new genus *Conglomerata*. *Journal of Invertebrate Pathology*, *159*, 95–104. https://doi.org/10.1016/j.jip.2018.10.003

Videvall, E. (2018). *Plasmodium* parasites of birds have the most AT-rich genes of eukaryotes. *Microbial Genomics*, *4*(2), e000150. https://doi.org/10.1099/mgen.0.000150

Wadi, L., & Reinke, A. W. (2020). Evolution of microsporidia: An extremely successful group of eukaryotic intracellular parasites. *PLOS Pathogens*, *16*(2), e1008276. https://doi.org/10.1371/journal.ppat.1008276

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Weiss, L. M., & Becnel, J. J. (2014). *Microsporidia: Pathogens of opportunity*. Wiley. https://doi.org/10.1002/9781118395264

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

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

Wiredu Boakye, D., Jaroenlak, P., Prachumwat, A., Williams, T. A., Bateman, K. S., Itsathitphaisarn, O., Sritunyalucksana, K., Paszkiewicz, K. H., Moore, K. A., Stentiford, G. D., & Williams, B. A. P. (2017). Decay of the glycolytic pathway and adaptation to intranuclear parasitism within Enterocytozoonidae microsporidia. *Environmental Microbiology*, *19*(5), 2077–2089. https://doi.org/10.1111/1462-2920.13734

Xiong, X., Geden, C. J., Bergstralh, D. T., White, R. L., Werren, J. H., & Wang, X. (2023). New insights into the genome and transmission of the microsporidian pathogen *Nosema muscidifuracis*. *Frontiers in Microbiology*, *14*. https://doi.org/10.3389/fmicb.2023.1152586

Ye, Z., Jiang, X., Pfrender, M. E., & Lynch, M. (2021). Genome-wide allele-specific expression in obligately asexual *Daphnia pulex* and the implications for the

genetic basis of asexuality. *Genome Biology and Evolution*, *13*(11), evab243. https://doi.org/10.1093/gbe/evab243

Žárský, V., Karnkowska, A., Boscaro, V., Trznadel, M., Whelan, T. A., Hiltunen-Thorén, M., Onut-Brännström, I., Abbott, C. L., Fast, N. M., Burki, F., & Keeling, P. J. (2023). Contrasting outcomes of genome reduction in mikrocytids and microsporidians. *BMC Biology*, *21*(1), 137. https://doi.org/10.1186/s12915-023-01635-w

Zhang, C., & Mirarab, S. (2022). Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. *Molecular Biology and Evolution*, *39*(12), msac215. https://doi.org/10.1093/molbev/msac215

Zimin, A. V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., Yorke, J. A., Dvořák, J., & Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research*, *27*(5), 787–792. https://doi.org/10.1101/gr.213405.116