Comparative genomics of microsporidia

Bryony A. P. Williams*[1], Tom A. Williams[2], Jahcub Trew[1]

[1]Biosciences, University of Exeter, Exeter EX4 4QD, UK

[2]School of Biological Sciences, University of Bristol, BS8 1TH Bristol, UK.

*Correspondence: b.a.p.williams @exeter.ac.uk.

Abstract:

The microsporidia are a phylum of intracellular parasites that represent the eukaryotic cell in a state of extreme reduction, with genomes and metabolic capabilities embodying eukaryotic cells in arguably their most streamlined state. Over the past 20 years, microsporidian genomics has become a rapidly expanding field starting with sequencing of the genome of *Encephalitozoon cuniculi* - one of the first ever sequenced eukaryotes, to the current situation where we have access to the data from over 30 genomes across 20+ genera. Reaching back further in evolutionary history, to the point where microsporidia diverged from other eukaryotic lineages, we now also have genomic data for some of the closest known relatives of the microsporidia such as *Rozella allomycis*, *Metchnikovella* spp. and *Amphiamblys sp.*. Data for these organisms allow us to better understand the genomic processes that shaped the emergence of the microsporidia as a group. These intensive genomic efforts have revealed some of the processes that have shaped microsporidian cells and genomes including patterns of genome expansions and contractions through gene gain and loss, whole genome duplication, differential patterns of invasion and purging of transposable elements. All these processes have been shown to occur across short and longer time scales to give rise to a phylum of parasites with dynamic genomes and a diversity of sizes and organisations.

Keywords genome compaction, microsporidia, genome evolution, gene loss, transposable elements

Why sequences genomes and where are we now?

Over the past decades genomic data have accumulated at a phenomenal rate with over 50 microsporidian genomes now available (Figure 1). This abundance of sequence data, curated at MicrosporidiaDB (Aurrecoechea et al. 2011), has provided a springboard for a multitude of cell and molecular biological work characterising the microsporidian cell structure and its interactions with hosts. In addition, these genomes represent an important data set for understanding the evolution of small eukaryotic genomes and parasite genomes more generally. An increasingly large amount of data in this single phylum allows us to get a better handle on the dynamics of reductive genome evolution in eukaryotes, understanding changes over smaller and larger evolutionary timescales in a group with large-scale variation in genome architecture and in which genome size varies over an order of magnitude.

The first sequenced microsporidian genome: *Encephalitozoon. cuniculi.*

The first genome of a microsporidian to be sequenced was that of *Encephalitozoon cuniculi* GB-M1 in 2001 (Katinka et al. 2001). Species within the clade *Encephalitozoon* emerged as agents of opportunistically infecting organisms in AIDS patients in the 1980s (Terada et al. 1987). Prior to the full sequencing of this organism's genome, full sequences of chromosomes had been decoded revealing the compact nature of microsporidian genomes (Duffieux et al. 1998; Peyret et al. 2001). Through preceding cell and molecular experiments it was also clear that microsporidia were unusually reduced eukaryotes that lacked the separate 5.8S rRNA gene distinctive of eukaryotes, had ribosomes that settled at a sedimentation coefficient equivalent to those extracted from bacteria (Ishihara and Hayashi 1968) and seemed to lack characteristic organelles such as the mitochondrion or a typical Golgi (Cavalier-Smith, Tom 1983). Whilst for some time leading up to the sequencing of the *E. cuniculi* genome these characteristics were viewed as primitive, there was growing evidence that microsporidia were related to fungi and that these reduced characteristics actually represented a secondary reduction (Cavalier-Smith, Tom 1983; Hirt et al. 1997, 1999; Roger 1999). However, it was with the full sequencing and annotation of the *E. cuniculi* genome that the extent and nature of reductive evolution within these organisms became clear (Katinka et al. 2001). At the time, the genome was completed by Sanger sequencing with the chromosome ends completely sequenced and annotated at a later date (Dia et al. 2016). The genome of *E. cuniculi* GB-M1 remains arguably the best annotated microsporidian genome and a key reference for all other microsporidia genomics studies.

Within the broader scope of microsporidian genome sizes, the *E. cuniculi* genome is small at 2.9 Mb, with few species having a smaller genome size. However, one characteristic of this genome which is common to all microsporidia is the small size of the predicted proteome, this currently stands at 2,041 predicted protein coding genes (The UniProt Consortium, 2021). The majority of these are single copy genes and the predicted proteins are short (359 amino acid mean length) relative to the average across eukaryotes (472 aa), but still longer that those in bacterial or archaeal genomes (320 aa and 283 aa respectively) (Katinka et al. 2001, Tiessen et al. 2012), which fits with the general trend of shorter proteins in genomes with small proteomes (Tiessen et al. 2012). These proteins have a simplified domain organisation with fewer domains on average than other eukaryotes (Koonin et al. 2004). It has been suggested that this results from a need for fewer interaction domains when there are fewer proteins to interact with (Zhang 2000; Koonin 2004). The genome is gene dense with a coding sequence for every 1.025 kb in the ‘chromosomal core’, just 38 introns and little repetitive DNA (Katinka et al. 2001; Pombert et al. 2013). The genome is organised as 11 linear chromosomes with a fairly homogeneous length ranging from 217 and 315 kb and is likely diploid (Brugère et al. 2000). Whilst remnant mitochondria named mitosomes have been uncovered in microsporidia (Williams et al. 2002), there is no mitochondrial genome, thus all the genomic material is contained within these linear chromosomes. The ends of the chromosomes are characterised by a SSU-LSU rDNA ‘transcription unit’ and areas of segmental repeats between 3.5 and 23.8 kbp. These segmental repeats can contain genes in multiple copies and also encode four different gene families (*interAE*, *interB*, *interC* and *interD*), hypothesised to mediate host-cell interaction and/or immune escape (Dia et al. 2007, 2016). Variation in the size of these telomeric areas is thought to be responsible for variation in size of chromosomes within the species (Biderre et al. 1999).

How do other microsporidian genomes differ from this model? The *E. cuniculi* genome size, content, and organisation reflect three key processes that have shaped the generally streamlined genomes of microsporidia. These are 1) loss of genes, 2) shortening of genes 3) loss of non-coding material including introns, intergenic regions repetitive DNA and transposable elements. Whilst there are mechanisms driving down genome size, within specific lineages or microsporidia, other mechanisms drive genome size in the opposite direction. These are the acquisition of new genes through gene and whole genome duplication, *de novo* gene origination, horizontal gene transfer, invasion and proliferation of transposable elements and re-expansion of intergenic regions. This interplay of mechanisms driving reduction and those driving genome expansion have led to a diversity of different genome sizes and organisations across the phylum where genome size varies by over 20-fold from 2.3 Mb to 53.1 Mb, whilst the number of predicted protein-coding genes varies over a much smaller range from approximately 1820 to 6,442 (Corradi et al. 2010; Cuomo et al. 2012; Pombert et al. 2015; Cormier et al. 2021). Over the next paragraphs we describe examples of how these different mechanisms have driven change in genome size, organisation and composition at different phylogenetic depths across the phylum (Figure 2).

Massive gene loss at the emergence of the microsporidia

Comparisons of gene family evolution in microsporidians and their fungal relatives have identified core gene families conserved across microsporidia as well as major changes in gene content that occurred early in the evolution of microsporidia, during adaptation to a host-associated lifestyle (Keeling et al. 2010; Cuomo et al. 2012; Heinz et al. 2012; Campbell et al. 2013; Nakjang et al. 2013; Desjardins et al. 2015; Wiredu Boakye et al. 2017; Galindo et al. 2018; Cormier et al. 2021). These studies identify a microsporidian ‘core’ set of protein families that are usually conserved across microsporidia. Some of these are ancestor-derived conserved gene families including genes associated with basic cellular functions such as transcription, translation and DNA replication and repair, cell cycle control, and the key enzymes for glycolysis, the pentose phosphate pathway, trehalose metabolism, chitin biosynthesis and those associated with the biosynthesis of structural components of the fungal cell membrane and spore wall (Nakjang et al. 2013). Proteins encoding a Fe/S cluster assembly pathway are also conserved across all sequenced microsporidian genomes (Freibert et al. 2017; Galindo et al. 2018). This pathway is localised to the mitosome and is likely essential for the maturation of certain nuclear and cytosolic Fe/S proteins (Freibert et al. 2017). The essentiality of this pathway provides a selection pressure for the retention of the mitochondrion in this highly minimal form across microsporidia.

In contrast some conserved microsporidia protein families are microsporidia-specific and are likely associated with, and essential to the ‘microsporidian lifestyle’. These include polar tube and spore wall protein families but also protein families without any characterised members or identifiable domains (Nakjang et al. 2013)

These studies also identified a massive loss of genes in the common ancestor of microsporidia involving the predicted loss of 1,590 protein families leaving a common ancestor of the ‘core’ microsporidia with an estimated 1121 protein families (Nakjang et al. 2013). This early gene bottleneck left microsporidia bereft of pathways associated with energy metabolism including oxidative phosphorylation and the tricarboxylic acid (TCA), likely allowing the degeneration of mitochondria. In addition to loss of energy metabolism, there is a severely limited repertoire for the biosynthesis of amino acids and no genes involved in the *de novo* synthesis of nucleotides. There is also generally thought to be a streamlining of the protein network of most core pathways and cellular components. Other canonical eukaryotic pathways are differentially lost and these include the RNA interference pathway, the spliceosome, fatty acid biosynthesis, the pentose phosphate pathway and glycolysis (Figure 2) (Akiyoshi et al. 2009; Keeling et al. 2010; Cuomo et al. 2012; Wiredu Boakye et al. 2017; de Albuquerque et al. 2020).

Most recent phylogenetic analysis suggest that microsporidia are closely related to the Fungi (James et al. 2013) but for a long time there was little genomic data for organisms that represent intermediate lineages between the microsporidia and other eukaryotes. However, there is now accumulating genomic data for many fascinating lineages that represent part of a sister group to the microsporidia or basal microsporidian lineages. These include, *Rozella allomycis*, *Amphiamblys Metchnikovella*, *Paramicrosporidium* and *Mitosporidium* (James et al. 2013; Mikhailov et al. 2017; Quandt et al. 2017; Galindo et al. 2018; Nassonova et al. 2021). *Rozella allomycis* is a Cryptomycotan that is an obligate parasite of the water fungus *Allomyces* and morphologically quite distinct from the microsporidia, in particular in that it has a flagellated stage, whilst microsporidia have lost the flagellum (James et al. 2006). *Amphiamblys*, *Metchnikovella*, *Paramicrosporidium* and *Mitosporidium* are morphologically more similar to the microsporidia and all possess a form of the polar filament or a manubrium analogous to the polar filament which is characteristic of microsporidia (Nassonova et al 2021*).* The relationship between these groups are unclear but phylogenetic analysis suggest that these organisms alongside microsporidia form a clade and that the microsporidia evolved from within a group of organisms called the Cryptomycota or Rozellomycota or Rozellida (Bass et al. 2018; Corsaro et al. 2019).

Extensive phylogenomic analyses have now shown how these lineages compare in gene content and organisation to those that have now been dubbed the 'canonical', 'long branch', or 'core' microsporidia giving some insight into events that shaped early microsporidia evolution (Bass et al. 2018; Timofeev et al. 2020; Nassonova et al. 2021). These studies have shown that the functional gene content of both *Amphiamblys* and *Metchnikovella* genomes is more similar to that of the 'core' microsporidia whilst that of *Paramicrosporidium* and *Mitosporidium* is more similar to *Rozella allomycis* and other eukaryotes, suggesting that much of the genome remodelling in the microsporidian lineage occurred after the divergence of these latter taxa (Figure 2). The genomes of *Paramicrosporidium* and *Mitosporidium* are relatively small and compact (7.28 Mb and 5.64 Mb respectively) (Haag et al. 2014; Mikhailov et al. 2017), but they nonetheless possess some metabolic characteristics that clearly set them apart from the microsporidia, such as mitochondrial genomes (Haag et al. 2014; Quandt et al. 2017). In contrast *Amphiamblys* and *Metchnikovella* are more metabolically reduced and like in 'core' microsporidia, mitochondrial pathways associated with the synthesis of nucleotides and amino acids are absent, and those associated with fatty acid metabolism are reduced (Galindo et al. 2018).

These data indicate that whilst genome compaction occurred sometime between the between the divergence of the Fungi and the Cryptomycota there were different rounds of gene loss that marked the transition from free living to parasitic organisms with the most dramatic loss in metabolic pathways likely occurring sometime before the divergence of *Amphiamblys* and *Metchnikovella* but after the divergence of the less metabolically simplified *Mi. daphniae*, *Pa. saccamoebae*, and *R. allomycis* (Galindo et al. 2018*).*

After these periods of massive gene loss, there is evidence of substantial gene family expansions in different microsporidian lineages, giving rise to microsporidia specific gene families(Nakjang et al. 2013). These extensive expansions have in particular affected transporter proteins, many of which are predicted to target the parasite cell membrane. The hypothesis is that these transporters allow import of metabolites from hosts filling the gaps created by the past large-scale loss of biosynthetic capacity (Nakjang et al. 2013). Two key examples of microsporidian and microspordian-lineage specific expansion of transporters include the nucleotide transporters NTT proteins that transport ATP and GTP and NAD+ from the host and the Major Facilitator Superfamily proteins (MFS) that transport ATP and GTP from the host. Both of these protein families are universally present across microsporidia, one having been acquired by horizontal gene transfer in the ancestor of the Cryptomycota + Microsporidia (see below) and the other being present in other eukaryotes with a canonical vertical inheritance from ancestors (Tsaousis et al. 2008; Cuomo et al. 2012; Major et al. 2019). Both have undergone multiple independent duplications within the phylum microsporidian lineages giving rise to different ortholog numbers in different lineages (Cuomo et al. 2012; Dean et al. 2018; Major et al. 2019).

In summary, the ancestor of ‘core or ‘long branch’ microsporidia was already a parasite that likely had a relatively compact genome and lacked a mitochondrial genome, the genes for the TCA cycle and oxidative phosphorylation, and de novo nucleotide and amino acid biosynthesis. However, it encoded proteins that allowed ATP uptake from the host, glycolysis, the pentose phosphate pathway, trehalose metabolism, RNAi and splicing.

Genomes changes over smaller time scales: Small largely static genomes in the genus *Encephalitozoon*: The genus *Encephalitozoon* harbours species with the smallest microsporidian genomes amongst this phylum of already reduced species and. Genomes from multiple species and multiple isolates of the same species have now been sequenced from within *Encephalitozoon* providing a glimpse into how these compacted genomes are evolving at the level of the genus. (Corradi et al. 2010; Pombert et al. 2012, 2013; Selman et al. 2013; Pelin et al. 2016). Multiple different strains of *E. cuniculi* named genotypes EC I-IV have been identified and are distinguished on the basis of the number of GTTT repeats within the internal transcribed spacer locus in the rRNA (Xiao et al. 2001). Representatives of genotypes, ECI, ECII, and ECIII have been fully sequenced and compared to with the originally sequenced *E. cuniculi* GB-M1 (ECI) strain to unveil patterns of change over a short evolutionary time scales (Pombert et al. 2012, 2013; Selman et al. 2013; Pelin et al. 2016). This revealed genomes with near identical gene contents and gene arrangements, however there was no evidence of recombination indicating that whist they are very closely related, they represent distinct strains with no exchange of material. In spite of their close relationship and highly similar genomes, this comparative study showed that the level of genetic diversity between strains was relatively high among ECI, ECII, and ECIII (4.2 SNPs/kb) relative to SNPs numbers among strains of other species of single celled pathogenic eukaryotes. This indicates a set of genomes that are largely frozen in content and organization but fast evolving at the level of the nucleotide and therefore individual genes (Pombert et al. 2013).

Whilst nucleotides are changing at a relatively fast rate, differences in the coding content or significant changes to coding sequences were very few: One gene encoding a protein with a homeobox domain present in ECI is absent from ECII/ECIII genomes, there is a gene fission in ECI relative to other strains, and the ECI strain encodes 3 similar paralogs of one protein whilst other strains have a single paralog (Pombert et al. 2013). However, the genome sequence of a wild isolate of genotype III from a steppe lemming did reveal a small scale change with a large potential impact on phenotype: A frameshift mutation in a key meiosis gene Spo1 which likely renders it incapable of meiosis and sexual reproduction (Pelin et al. 2016). This demonstrates that, despite very high levels of similarity in the genome, there do exist small differences between very closely related strains that may result in fundamental differences in biology.

Within the same genus, further *Encephalitozoon* species have been sequenced (Corradi et al. 2010; Pombert et al. 2012). Whilst the genome of *E. cuniculi* is already highly compact, *E. intestinalis* takes that reduction a step further: Its genome is just 2.3 Mb and thus 20% smaller than that of *E. cuniculi* (Corradi et al. 2010). Its genome is even more gene dense, its genes are even shorter and it has fewer hypothetical proteins and gene duplicates (Corradi et al. 2010). The two genomes share the same introns and are almost colinear in their gene order across their chromosomal cores, emphasising very few differences between the two genomes, but large blocks of the subtelomeric regions present in *E. cuniculi*. are absent in *E. intestinalis.* This results either from a process of expansion in *E. cuniculi* or further contraction in *E. intestinalis* relative to their common ancestor (Corradi et al. 2010). The genomes of *E. romaleae*, a grasshopper pathogen and *E. hellem,* another human pathogen, have also been sequenced (Pombert et al. 2012). Again, these show very high levels of genome reduction and an extremely similar gene content to *E. cuniuli* with the same structure of 11 linear chromosomes (Pombert et al. 2012). One striking observation, however, was the presence of a particular set of genes encoding pathways for folate salvage, *de novo* folate biosynthesis, and purine metabolism. Components of these pathways are represented by genes across four chromosomal regions and are absent from the other sequenced *Encephalitozoon* genomes (Pombert et al. 2012). These are pathways that are not typically complete in other sequenced microsporidia species and phylogenetic analysis suggest that these genes were acquired by horizontal (or lateral) gene transfer, but interestingly not from a single donor organism source. Rather this pathway has been cobbled together by an ancestor of *E. hellem* and *E. romaleae* using genes from different sources, apparently including several prokaryotes and, for one gene, an animal or fungal source (Pombert et al. 2012). The genes exist in the subtelomeric regions of these genomes which, as seen in *E. cuniculi* are the sites of rapid change and recombination that are home to multicopy protein families and pseudogenes, and insertion of a horizontally acquired gene here it is far less likely to cause a disruption than in the extremely dense 'chromosomal cores'. Curiously, whilst the pathways appear intact in *E. hellem,* they look to be in the process of degeneration in *E. romaleae* (Pombert et al. 2012).

Looking just outside the genus, the genome of close relative *Ordospora colligata,* a pathogen of *Daphnia*, has also been sequenced. This shows a very similar genome content and organisation with a small compact genome (1820 predicted coding sequences within a 2.3 Mb assembly), with little repetitive DNA, a handful of introns, but organised into 10 chromosomes, rather than the 11 seen in *Encephalitozoon* spp.). Like the genomes of *E. hellem* and *romaleae*, this genome has been shaped by horizontal gene transfer. In this case, the *Daphnia* host is putatively the source for a

Septin gene within the *O. colligata* genome. This Septin shares structural features with Septin 7 which in the pathogen *Candida albicans* is an effector that induces the uptake of the pathogen by the host. Septin 7 is speculated to have an analogous function in *O. colligata*, allowing an alternative means of entry into the host cell than via the polar tube (Pombert et al. 2015). These examples of pathway acquisition through horizontal gene transfer (and others below) demonstrate a mechanism of expansion of the metabolic repertoire after the initial large-scale gene and pathway loss in the ancestor of the microsporidia.

Horizontal gene transfer driving innovation and change in microsporidian genomes:

The extent to which eukaryotic genomes are shaped by horizontal gene transfer is a subject of active debate (Ku and Martin 2016; Martin 2017; Leger et al. 2018; Van Etten and Bhattacharya 2020). Within the microsporidia horizontal gene transfer has unquestionably been important in driving at least one evolutionary transition within the phylum. This key horizontal gene transfer was the acquisition of nucleotide transport proteins (NTT). This acquisition likely occurred from bacteria into the common ancestor of Microsporidia and *Rozella* spp (Dean et al. 2018). This was followed by multiple duplication events within the microsporidian lineage give rise to multiple copies (often 4) of this protein within microsporidia genomes (Dean et al. 2018). As proteins that allow the acquisition of host ATP, GTP and NAD+, this horizontal gene transfer event led to the evolutionary transformation of the microsporidian lineage into energy parasites and likely allowed the loss of energy metabolism pathways and the degeneration of the mitochondrion. More broadly it appears that horizontal gene transfer has been an important force in shaping microsporidia genomes content, with apparently up to 2.2% of microsporidian genes derived from horizontal gene transfer (Alexander et al. 2016). It is likely the intracellular lifestyle and the intimate association between parasite and host that has facilitated the transfer of DNA and genes between animals and microsporidia perhaps through integration of reverse transcribed host mRNA into the microsporidian genome (Alexander et al. 2016). Examples of host to microsporidia transfer include multiple transposable elements (see below), and the septin and a purine nucleotide phosphorylase (PNP) mentioned above (Selman et al.

2011; Pan et al. 2013; Parisot et al. 2014; Pombert et al. 2015). However, intracellular life is also hypothesised to drive horizontal gene transfer from prokaryotes into microsporidia (Campbell et al. 2013; Alexander et al. 2016). One way in which this may occur is through contact with bacterial DNA within host phagocytic vesicles (Alexander et al. 2016). Coinfection of a host cell alongside intracellular bacterial pathogens such as *Chlamydia* spp. has also been suggested to facilitate prokaryote to microsporidian horizontal gene transfer (Lee and Heitman 2017). This scenario might have allowed the acquisition of the NTT transporters by microsporidia from *Chlamydia* spp,. which are also energy parasites and one of the bacterial lineages whose genomes encode homologs of these NTT proteins (Major et al. 2017).

Small genomes with no introns: The genus *Nematocida*

Among the highly-derived, “long branch” or “core” microsporidians, the deepest phylogenetic split appears to lie between the genus *Nematocida* and the rest of long-branch microsporidian diversity (Figure 2). Multiple *Nematocida* species genomes have now been sequenced, and comparisons between these and other microsporidians can therefore provide insight into the early evolution of the group following the transition to an obligate intracellular parasitic lifestyle. Species within this genus, as the name implies, infect nematodes and despite the large phylogenetic distance between this genus and *Encephalitozoon* the two clades share some general genome characteristics. Within this genus we observe another example of very small and compacted genomes ranging in assembly size from 3 to 4.4 Mb. For example, the *Nematocida parisii* ERTm1 genome is ~4.1 Mb in size and predicted to be comprised of 72.8% coding sequence with a mean distance of 418 bp between genes (Cuomo et al. 2012). *Nematicida* genomes also show the same level of reduced metabolic capacity as species within the genus *Encephalitozoon*, except for presence of a CTP synthase which would allow *Nematicida* spp. to synthesise CTP from UTP (Cuomo et al. 2012). Within the genus there exists some size variation in genomes over a relatively short evolutionary timescale. The earliest diverging known *Nematocida* species*, Ne. displodere* has a more reduced genome, similar to species of *Encephalitozoon* in genome size with a higher proportion of coding material (85.8% compared to 69.2% for *Ne*. *parisii* and 63.7% for *Nematocida ausubeli*) and fewer proteins (2278 compared to 2661 in *Nematocida parisii*) (Cuomo et al. 2012; Luallen et al. 2016). However, interestingly, *Nematocida displodere* has a considerable proportion of its genome occupied by members of a single large, expanded gene family . Whilst all sequenced *Nematocida* genomes house expanded gene families (Reinke et al. 2017), the *Nematocida* large gene family 2 has 235 members in *Ne. displodere*, and contributes over 10% of its predicted proteome (Luallen et al. 2016). Many of the protein products (152/235) have predicted signal sequences and/or a RING domain (113/235) and like the *interAE*, *interB*, *interC* and *interD* gene families in *Encephalitozoon*, these genes are thought to encode proteins that mediate interactions with hosts (Dia et al. 2007; Reinke et al. 2017). Similar features are found in gene families across many other lineages of microsporidia (see below), and expanded gene families within the genus *Nematocida* where they represent a substantial proportion of identified ‘host exposed proteins’ (Reinke et al. 2017). These ‘host exposed proteins’ are parasite proteins that have been identified within the host cell cytoplasm or nucleus using a method called spatially restricted enzymatic tagging (Reinke et al. 2017).

Whilst *Nematocida* species share the same reduced metabolic capability as *Encephalitozoon*, one key difference between *Nematocida* and *Encephalitozoon* genomes is that in *Nematocida* spp., no introns were detected in gene sequences, there was no evidence of spliced transcripts, and in addition many components of the spliceosomal machinery are lost (Cuomo et al. 2012).

Independent losses of splicing and introns:

Although introns are sparse within the microsporidia a small number have been found in most microsporidian genomes (Whelan et al. 2019). These introns appear to fall into two broad categories. The first includes short introns of around 25 bp which typically occur in ribosomal proteins right at the start of the coding region (typically after the first ATG). The second category includes two specific introns found in the same two genes across different species of microsporidia (Whelan et al. 2019). The longer introns are almost always found to have been removed by splicing (in transcriptomes 80% of the time) whereas the short introns are rarely removed (splicing rates are as low at 20%) (Whelan et al. 2019).

On at least three independent occasions, introns and the splicing machinery have been entirely lost during microsporidian evolution: Once in the *Nematocida* clade, once in the *Vittaforma*/Enterocytozoonidae clade and once in *Edhazardia aedis* (Figure 2) (Akiyoshi et al. 2009; Cuomo et al. 2012; Wiredu Boakye et al. 2017). This is supported by both an absence of observable introns and a loss of many genes encoding components of the spliceosome from these genomes (Akiyoshi et al. 2009; Cuomo et al. 2012; Wiredu Boakye et al. 2017; Whelan et al. 2019). This pattern of retention and loss of introns and splicing is hard to explain. Introns in microsporidia typically fall in ribosomal proteins which are obviously fundamental to cell function. It has been suggested that there may be a role for these shorts introns in regulating their own expression and therefore ribosome synthesis as is seen in *S. cerevisiae* (Roy et al. 2020), yet why they are retained is some species but not others, remains unresolved.

Larger genomes with polyploidisation and transposable elements: *Nosema/Vairimorpha* spp.

The closest relatives to the *Ordospora*/*Encephalitozoon* clade with sequenced genomes are those in the *Nosema*/*Vairimorpha* group. Here, again there exists extensive genome data both within genera and within a single species and this shows an upwards shift in genome size relative to *Ordospora*/*Encephalitozoon* and the *Nematocida* genomes. The genus *Nosema* has been somewhat problematic in phylogenetic placement in the past as it has included multiple species that have since been transferred to different groups within the phylogenetic tree of microsporidia (Sokolova et al. 2005; Lord et al. 2010). In addition, *Nosema* and *Vairimorpha* species did not separate into clades and were intermixed in molecular phylogeny. This group contains species isolated from arthropods, mainly insects and particularly Lepidoptera and Hymenoptera (Tokarev et al. 2020). Recently, however these clades have been redefined on the basis of phylogenetic position with the key bee pathogens *Nosema apis* and *ceranae* being redefined as *Vairimorpha* (Tokarev et al. 2020). Nevertheless, these two genera together form a monophyletic clade of closely related pathogens that infect arthropods. As a pathogen that has been implicated in poor honeybee health *Vairimorpha ceranae* has been of intense interest to the insect pathology community. and genomes of multiple isolates of *V. ceranae* have now been sequenced (Cornman et al. 2009; Pelin et al. 2015; Peters et al. 2019; Huang et al. 2021).

The first genome sequence of *V. ceranae* revealed an estimated genome size of 7.86 Mbp and a total of 2,641 putative protein-coding genes and whilst larger than those of *Encephalitozoon* species, it is less gene dense, with 0.60 genes/kb (64.8% coding sequence) and contains more repetitive DNA, particularly transposable elements (Cornman et al. 2009). These transposable elements are diverse including members of the gypsy-type LTR retrotransposons and DNA transposons such as Merlin, Helitron, piggyBac, and MULE and occupy over 20% of the genome (Cornman et al. 2009; de Albuquerque et al. 2020). *V. ceranae* infects *Apis melifera* worldwide but was originally described in the Asiatic honeybee (*Apis cerana*) and extensive genome sequencing has been used to better understand the origins of this pathogen and its global spread within bee populations (Pelin et al. 2015; Peters et al. 2019). SNP phylogenies inferred from globally distributed isolates demonstrated a lack of geographic structure, suggesting a recent spread of the pathogen among hives, perhaps as a result of human activity (Pelin et al. 2015).

This study also revealed a surprising level of within-individual genetic variation in global *V. ceranae* isolates that might be explained by a polyploid (at least tetraploid) genome in comparison to the ancestral diploid population in Asia (Pelin et al. 2015; Peters et al. 2019). Within the genomes both of the Asian and global populations there are high of levels of linkage disequilibrium pointing to a likely clonal life style and a lack of recombination (Pelin et al. 2015; Peters et al. 2019).

V. ceranae forms a clade with two other species that have sequenced genomes, one also infecting honeybees, *V. apis* and one infecting lepidopterans, *Nosema* YnPr (Figure 2 and Tokarev et al. 2020) although the evolution of host preference within the clade remains unclear. The genome of *No.* YnPr is smaller at 3.4 Mb and is more compact with shorter genes, whilst that of *V. apis* is larger at 8.5 Mb (Chen et al. 2013; Xu et al. 2016). A major contributor to this difference in size is a difference in the number and diversity of transposable elements in the two genomes, with the larger genome having a greater proportion of transposable elements, suggesting that these can invade and/or expand within genomes rapidly over relatively short evolutionary time periods within the microsporidia.

Transposable element expansion is perhaps more obvious in the true *Nosema* clade in which the silkworm pathogens *No. antheraeae*, *No. bombycis* and the gammarid pathogen *No. granulosis* have genome assembly sizes of 6.6 Mb, 15.7 Mb and of 8.8 Mb respectively (Pan et al. 2013; Cormier et al. 2021). In fact, *No. bombycis* is reported to have the largest proportion of transposable element content of any microsporidia genome sequenced so far (Pan et al. 2013; de Albuquerque et al. 2020). This has likely resulted from the expansion of transposable elements common across microsporidia such as Ty3/Gypsy LTR retrotransposons, and also through acquisition of new elements from insect hosts, particularly the relatively recent acquisition of multiple *Piggybac* elements from lepidopteran hosts possibly in the ancestor of *No. antheraeae* and *No. bombycis*, followed by more extensive proliferation in the genome of *No. bombycis* (Pan et al. 2013) .

Enterocytozoonidae: Metabolic reductionism to the extreme.

The Enterocytozoonidae clade is home to some of the economically most important microsporidia species, for example *Enterocytozoon bieneusi,* the microsporidian species most commonly found in humans (Didier and Weiss 2006) and *Enterocytozoon hepatopenaei* (dubbed EHP), a microsporidian that is currently causing extensive damage to the shrimp industry (Thitamadee et al. 2016). Several genomes within this family have been sequenced and whilst they are not the smallest from the perspective of number of nucleotides (3.1 - 6 Mb), they are unique with respect to the extent of loss of metabolic pathways (Akiyoshi et al. 2009; Keeling et al. 2010; Wiredu Boakye et al. 2017). The loss of the ability to generate ATP in mitochondria is a key characteristic of the microsporidia and whilst the intracellular milieu allows access to ATP via transporters, it is thought that glycolysis is crucial in the extracellular spore stage when host resources are inaccessible (Dolgikh et al. 2011; Heinz et al. 2012; Timofeev et al. 2020). In the spore stage, the glucose to generate ATP through glycolysis likely comes from the breakdown of trehalose (Timofeev et al. 2020). Thus, the enzymes of the glycolytic pathway and trehalose metabolism are considered core to microsporidian energy metabolism (Nakjang et al. 2013).

However, the genomes sequenced from the Enterocytozoonidae and that of the closely related *Hepatospora eriocheir* do not encode full glycolytic pathways, but partial pathways with several components lost which likely render these organisms incapable of glycolysis (Akiyoshi et al. 2009; Keeling et al. 2010; Wiredu Boakye et al. 2017). Additionally these species have lost the trehalase enzyme that could potentially break down trehalose into glucose in the spore, and key components of the pentose phosphate pathway such as transketolase (Keeling et al. 2010; Wiredu Boakye et al. 2017). These losses seem to leave these species apparently entirely dependent on their hosts for ATP. Whist there has been little research on the physiological state of microsporidian spores, it has been suggested that the process of rapid polar tube extrusion during spore germination must be ATP dependent (Timofeev et al. 2020). However, the absence of ATP generating pathways in these taxa suggests that either it is not, or that these particular species enter the host by another mechanism, potentially phagocytosis (Wiredu Boakye et al. 2017).

In addition to the loss of these core energy metabolic pathways, these species have also far fewer proteins associated with fatty acid biosynthesis in comparison to other microsporidia. Microsporidia are generally limited in their repertoire of fatty acid biosynthesis enzymes and elegant experiments to knock down host lipid metabolic processes in *Tubulinosema ratisbonensis* infected *Drosophila* have started to unpick the dependency of microsporidia on host lipids (Franchet et al. 2019). However, within the Enterocytozoonidae there are further losses of enzymes associated with glycerophospholipid metabolism and specifically the generation of phosphatidylethanolamine and phosphatidylcholine (Keeling et al. 2010; Wiredu Boakye et al. 2017). These enzymes are otherwise considered broadly conserved within the microsporidia and their loss in the Enterocytozoonidae suggests a potential dependence of the host for these key components of biological membranes (Nakjang et al. 2013; Wiredu Boakye et al. 2017). In addition to these losses of key metabolic pathways, there are no introns in these genomes nor do they encode all the genes necessary to form the spliceosome (Akiyoshi et al. 2009; Keeling et al. 2010; Wiredu Boakye et al. 2017).

Species with larger proteomes: what accounts for the differences?

With the publication of the *Trachipleistophora hominis*, *Nosema ceranae, Hamiltosporidium tvaerminnensis* and *Enterocytozoon bieneusi* genomes it became clear that some species of microsporidia had larger proteomes than others (Corradi et al. 2009; Akiyoshi et al. 2009; Cornman et al. 2009; Heinz et al. 2012)**.** *Trachipleistophora hominis* for example with a genome size of ~11.6 Mb has a predicted proteome that is approximately 30% larger than those of small-genome microsporidia (Heinz et al. 2012; Watson et al. 2015). More recently, the genome sequence of the *Gammarus*-infecting species *Dictyocoela muelleri* was published with 6,442 predicted genes (Cormier et al. 2021). Whilst there are some differences in the complement of metabolic pathways between microsporidia (as seen above), larger proteome size is not associated with a greater metabolic capacity. In those microsporidia that have larger proteomes, the vast majority of extra proteins have no annotated function and many are result of expansion of microsporidian or species-specific gene families through gene duplication (Heinz et al. 2012; Nakjang et al. 2013; Pan et al. 2013; Cormier et al. 2021). For example the 2,591 of 6,442 genes predicted in the *D. muelleri* genome can be clustered in just 233 orthogroups (Cormier et al. 2021). The genome of *Nosema bombycis*, the pathogen of the domesticated silkworm, encodes 4,458 predicted proteins relative to 3,413 in the genome of the close relative *No. antheraeae* which infects the Tussar silk moth (Pan et al. 2013) and many of the extra genes in *No. bombycis* have arisen as tandem repeats since the divergence of *No. bombycis* and *No. antheraeae* (Pan et al. 2013). As seen in *Nematocida*, one or a handful of gene families can account for a considerable number of predicted proteins and this is also the case in many other microsporidian genomes (Reinke et al. 2017). These extreme gene family expansion events can substantially add to the predicted proteome and has led to both species-specific expansions such as the '*Nematocida* expanded gene family 1' proteins or the more widespread InterB proteins found in *Encephalitozoon*, *Vittaforma*, and *Anncaliia* (Dia et al. 2007). These families can have over 100 members, are often enriched with motifs that mediate protein-protein interactions and many of them are predicted to have secretion signals that may direct them out of, or to the surface of the microsporidian cell (Heinz et al. 2012; Campbell et al. 2013; Reinke et al. 2017). The functional importance of these expanded protein families is not fully understood but they are strong candidates for characterisation as potential secreted parasite effector proteins.

In other instances, proteome increase has come about through whole genome duplication. Within the genus *Spraguea* there is evidence of a recent whole genome duplication in the ancestor of the North American population of *Spraguea* inevitably leading a duplicate copy of each gene (Williams et al. 2016). Many of the resultant duplicates are pseudogenes but some are retained as intact open reading frames. As a result, the lineage that has experienced this whole genome duplication has a higher predicted protein number than close relatives (Williams et al. 2016). Novel lineage specific genes can also arise *do novo* from previously non-coding material rather than via gene duplication or rearrangement of exons/introns or other genomic elements. These *de novo* genes can be identified as open reading frames that are unique to one species and are actively transcribed, but where there is clear sequence similarity at the DNA level to non-coding regions from close relatives (Figure 3) (Nakjang et al. 2013; Williams et al. 2016).

Genome inflation through transposable element bursts and intergenic region expansion: *Edhazardia aedis* and *Anncaliia algerae.*

Differences in genome organisation and content, such as smaller or larger gaps between genes or the presence or absence of certain genes or gene families can account for some of the difference between genome sizes, but large genome differences over short times scale are sometimes driven by bursts of transposable elements (Naito et al. 2006). There is a general correlation between genome size and transposable element content in eukaryotes broadly (Elliott and Gregory 2015) and this same trend applies broadly to the microsporidia: The smallest genomes have no or very few transposable elements, while the largest harbour large and diverse transposable element populations (Parisot et al. 2014).

Genome size expansion through the acquisition and spread of transposable elements has occurred multiple times within the microsporidia. One example seen above is in the genome of *Nosema bombycis* compared to the genomes of its close relatives (Pan et al. 2013). However, transposable elements are found in genomes across the diversity of microsporidia and have been shown to include both non-LTR retrotransposons and LTR retrotransposons and DNA transposons (including Helitron, Mariner, Tc1, Merlin and piggyBac). One genome that has become particularly expanded by a transposable element burst is that of the species *Anncaliia algerae* (Parisot et al. 2014). This pathogen primarily infects insects but is also a serious opportunistic pathogen of humans (Coyle et al. 2004). With a genome size of 23 Mb, it has one of the larger microsporidian genomes and in this species there a large diversity of transposable elements with 97 LTR retrotransposons, four non-LTR retrotransposons, and 139 DNA transposons identified within a single *An. algerae* genome (Parisot et al. 2014). Many of the transposable elements seen in microsporidia likely originate through horizontal gene transfer from hosts permitted by the intimate association between microsporidian and their host organism. Interestingly, recent work has shown that there are substantial differences in transposable element content between different strains of *Anncaliia algerae*, indicative of a recent change in of transposable element content driven primarily by the spread of DNA transposons (de Albuquerque et al. 2020).

It has been suggested that the RNAi pathway in microsporidia has a role in transposable element defence and the retention or loss of the pathway is driven by the presence or absence of transposable elements within microsporidian genomes (Nakjang et al. 2013) and whilst there is an association between the retention of Dicer and Argonaute and the accumulation of transposable elements in microsporidian genomes, there is not a complete match and other forces such as drift may also play a role in dictating when are where they are retained across the phylum (de Albuquerque et al. 2020).

Whilst larger microsporidian genomes have often become swollen through acquisition of transposable elements, one genome that bucks this trend is that of the mosquito parasite *Edhazardia aedis* which has the largest genome of any sequenced microsporidian to date at 51.3 Mb (Desjardins et al. 2015). This genome, however, is not characterised by more transposable elements: Only 5.5% of the *Edhazardia aedis* genome could be classified as repetitive, and – in common with some smaller microsporidian genomes – it entirely lacks introns (Desjardins et al. 2015). *Edhazardia aedis* encodes 4190 protein-coding genes, a relatively large complement for a microsporidian but not making up the difference in genome size, with only 9% of the genome coding for protein. Intergenic distance tends to increase with increased genome size in the microsporidia (Figure 4), but this is particularly apparent in the genome of *Edhazardia aedis* which has become bloated to it large size primarily through the accumulation of expanded AT-rich intergenic regions (Desjardins et al. 2015). These can be 10s of thousands of bases in length (Desjardins et al. 2015) and are speculated to allow additional regulation of gene expression (Troemel and Becnel 2015). Phylogenetically, *Edhazardia aedis* emerges from within a clade of small-genome microsporidians, suggesting that its larger genome is the result of secondary re-expansion rather, than the retention of the larger intergenic regions more typical of other eukaryotes.

What drives the evolution of microsporidian genome size?

Adaptive or functional explanations for diminutive microsporidian genome sizes have been invoked in the past (Cavalier-Smith 2005). Three factors have been implicated in driving down genome size. These first is metabolic reduction, explained by loss of selective pressure to retain certain pathways after the adoption of an intracellular lifestyle as the host allows access to these resources. The second is economy of resources required to replicate DNA in the nucleus as a force driving miniaturisation. The third is spatial reduction with the small cell size driving decreased genome size to maintain a ‘karyoplasmic ratio” (Cavalier-Smith 2005). In contrast it has also been proposed that larger genomes with more genes are the result of selection for a large gene repertoire in species that have more than one host type (Peyretaillade et al. 2012; Pan et al. 2013)

However, these functional explanations, as in other eukaryotes, do not provide a complete account of differences in genome sizes and organisations over short time scales, or vast differences in genome sizes between organisms with very similar hosts and fundamentally the same intracellular niche. In addition to selection acting on genome size changes, genome size variation can also be driven by neutral processes. Whether genomic changes such as point mutations; gene duplications, transfers or losses; or acquisition of transposable elements will become fixed within a lineage depends largely on the fitness cost or benefit of that change. In populations with large effective population sizes natural selection will allow changes with a small benefit to be fixed and mutations with small costs to fitness will be purged (Ohta 1992; Lynch and Conery 2003). However, in organisms with small effective population sizes, natural selection is less effective and slightly deleterious mutations such as the addition of a new non-coding genomic element may become fixed through drift (Lynch and Conery 2003).

Non-adaptive processes in shaping microsporidian genomes have been recently investigated using genomic data (de Albuquerque et al. 2020; Haag et al. 2020). Firstly, in order to investigate how genome architecture has evolved in distantly related microsporidian species with similar hosts but different life histories Haag *et al* sequenced seven new microsporidian genomes infecting *Daphnia magna*: *H. tvaerminnensis* (x 2 strains) and *H. magnivora* (x 2 strains), and *Ordospora colligata* (x 4 strains) (Haag et al. 2020). These organisms were chosen because they inhabit the same host, but they display a mix of sexual and asexual lifestyles and patterns of horizontal transmission or vertical transmission (or even a mix of both within the same species) allowing the impact of these differences to be evaluated. These species also differ vastly in genome size and organisation. *Ordospora* is strictly horizontally transmitted and has a compact genome of 2.3 Mb in comparison to the genomes of *Hamiltosporidium* spp. which range from 17.2 Mb to 25.2 Mb and are a mixture of vertically and horizontally transmitted (Haag et al. 2020). The *Hamiltosporidium* genomes, like many expanded microsporidian genomes are characterised by longer intergenic regions, segmental duplications and a greater proportion of transposable elements. One explanation for the differences in architecture between these groups is that vertical transmission creates a population bottleneck decreasing effective population size and thus decreasing the power of selection to remove slightly deleterious mutations (Haag et al. 2020). This is borne out by the observation that those species with larger genomes and vertical transmission as a part of their life stage also have an excess of nonsynonymous substitutions in single-copy orthologous genes relative to horizontally transmitted species (Haag et al. 2020).

Building on this hypothesis, Alberquerque et al looked at the distribution of transposable elements across the phylum studying the pattern of where transposable elements occur and where they have expanded within genomes (de Albuquerque et al. 2020). Comparing 47 microsporidian genomes confirms the positive relationship between genome size and transposable element content that exists in eukaryotes generally (Elliott and Gregory 2015). It also showed statistically significant differences in the percentage of the genome taken up by transposable elements and genome size generally in those species with vertical transmission versus those that only transmit horizontally where species with vertical transmission have a higher proportion of transposable elements, and larger genomes (de Albuquerque et al. 2020). This reinforces the idea that vertical transmission is associated with genome expansions and the spread of transposable elements in microsporidia, and this difference is the result of genetic drift in populations with a small effective size rather than selection (de Albuquerque et al. 2020).

The power, limitations and future of genomic studies in microsporidia:

Microsporidian genomes are small, and these organisms evolve under very different selective constraints compared to free-living eukaryotes. Nonetheless, genomic studies of microsporidia have demonstrated that, particularly in regard to evolutionary dynamics of genome size change, the group represent a microcosm of evolutionary processes occurring within eukaryotes more generally.

Now comparative genomic studies are not only describing the differences in organisation and content and the mechanisms that led to these changes, but also beginning to provide some answers as to which evolutionary pressures drive the differences in microsporidian genome size.

Genomic studies have given real insight into nature of reductionism and patterns of differential reduction of metabolic and cellular processes during the evolution of the phylum. This has revealed clear patterns of pathways that are universally lost across the phylum for example the electron transport chain and pathways that are universally conserved across the microsporidia (or example mitochondrial iron sulphur cluster assembly proteins in the mitosome. In addition, genomic studies have highlighted pathways and complexes that have degenerated at certain points during the evolution of the phylum (RNAi, glycolysis and the spliceosome) and the general paring down of eukaryotic complexes to something far more minimal than seen in ‘typical’ eukaryotes.

In terms of understanding cellular processes, genome data can suggest hypotheses that can be tested by experiment. For example, comparative genomic studies can generate lists of candidate ORFs associated within particular unusual lifestyles for example feminisation of hosts or occupation of a intranuclear niche (Wiredu Boakye et al. 2017; Cormier et al. 2021). However, the lack of a genetic system for microsporidia continues to impede progress in understanding the biological significance of these ORFs. Nonetheless, innovative experiments informed by genomic observations continue to provide new insights into microsporidian biology and the biology of parasitism, as illustrated for example by recent work on the evolution of the microsporidian ribosome (Barandun et al. 2019; Ehrenbolger et al. 2020). The ribosome was one of the earliest noted examples of reduction in the microsporidia and when the *E. cuniculi* genome was first sequenced, 70 different ribosomal proteins were identified in contrast to the ~80 typically found in other eukaryotes, suggesting the loss of several proteins (Katinka et al. 2001). In addition, the ribosomal RNA genes in microsporidia are extensively shortened so that the eukaryote-specific expansion segments that mediate many interactions with proteins are lost and it was therefore suggested that several ribosomal proteins are no longer part of the ribosome but have extra-ribosomal functions (Melnikov et al. 2018). However recent generation of cryo-EM structures and proteomic characterisation of the *Vairimorpha necatrix* and *Paranosema locustae* ribosomes give an in depth insight into of the nature of ribosomal reduction (Barandun et al. 2019; Ehrenbolger et al. 2020). These studies demonstrated that, in fact, some microsporidia have retained all the ribosomal proteins typically present in fungi apart from eL38 and eL41. Whilst expansion segments are lost from rRNA, this loss is compensated for proteins being held in place by protein-protein interactions. This illustrates that whilst reduction of the ribosome has occurred through shortening of genes and proteins, proteins are retained and the core structure of the eukaryotic ribosome is preserved. This study highlights just how challenging it is to make predictions about the biology of microsporidia on the basis of bioinformatic analysis and whilst this is true of studies in all organisms, it is particularly pertinent in the microsporidia. This is partly because the very high rates of evolutionary change in proteins sequences means that it is not always easy to identify homologs of short or particularly divergent proteins, and because the process of reduction in microsporidia has resulted led to unique solutions, and highly derived states, which cannot be predicted fully without functional data.

However, the available genomes represent an essential resource to support in depth studies like those of the ribosome and as new emerging technologies are applied to this fascinating group of organisms it will undoubtedly reveal novel ways in which microsporidia have adapted eukaryotic complexes in a way that cannot be understood with genomic data and bioinformatics alone.

References:

Akiyoshi DE, Morrison HG, Lei S, et al (2009) Genomic survey of the non-cultivatable opportunistic human pathogen, *Enterocytozoon bieneusi*. PLoS Pathog 5:e1000261. https://doi.org/10.1371/journal.ppat.1000261

Alexander WG, Wisecaver JH, Rokas A, Hittinger CT (2016) Horizontally acquired genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides. Proc Natl Acad Sci U S A 113:4116–4121. https://doi.org/10.1073/pnas.1517242113

Aurrecoechea C, Barreto A, Brestelli J, et al (2011) AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. Nucleic Acids Res 39:D612-619. https://doi.org/10.1093/nar/gkq1006

Barandun J, Hunziker M, Vossbrinck CR, Klinge S (2019) Evolutionary compaction and adaptation visualized by the structure of the dormant microsporidian ribosome. Nat Microbiol 4:1798–1804. https://doi.org/10.1038/s41564-019-0514-6

Bass D, Czech L, Williams BAP, et al (2018) Clarifying the Relationships between Microsporidia and Cryptomycota. Journal of Eukaryotic Microbiology 65:. https://doi.org/10.1111/jeu.12519

Biderre C, Mathis A, Deplazes P, et al (1999) Molecular karyotype diversity in the microsporidian *Encephalitozoon cuniculi.* Parasitology 118 (Pt 5):439–445. https://doi.org/10.1017/s0031182099004023

Brugère JF, Cornillot E, Méténier G, et al (2000) *Encephalitozoon cuniculi* (Microspora) genome: physical map and evidence for telomere-associated rDNA units on all chromosomes. Nucleic Acids Res 28:2026–2033. https://doi.org/10.1093/nar/28.10.2026

Campbell SE, Williams TA, Yousuf A, et al (2013) The Genome of *Spraguea lophii* and the Basis of Host-Microsporidian Interactions. PLoS Genetics 9:. https://doi.org/10.1371/journal.pgen.1003676

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

Cavalier-Smith, Tom (1983) A 6-kingdom classification and unified phylogeny. In: W. Schwemmler and H. E. A. Schenk, eds., Endocytobiology II: intracellular space as an oligogenetic ecosystem. De Gruyter, Berlin, pp 265–280

Cavalier-Smith, Tom (2005) Economy, Speed and Size Matter: Evolutionary Forces Driving Nuclear Genome Miniaturization and Expansion. Annals of Botany 95:147–175. https://doi.org/10.1093/aob/mci010

Chen YP, Pettis JS, Zhao Y, et al (2013) Genome sequencing and comparative genomics of honey bee microsporidia, *Nosema apis* reveal novel insights into host-parasite interactions. BMC Genomics 14:451–451

The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49:D480–D489. https://doi.org/10.1093/nar/gkaa1100

Cormier A, Chebbi MA, Giraud I, et al (2021) Comparative Genomics of Strictly Vertically Transmitted, Feminizing Microsporidia Endosymbionts of Amphipod Crustaceans. Genome Biol Evol 13:. https://doi.org/10.1093/gbe/evaa245

Cornman RS, Chen YP, Schatz MC, et al (2009) Genomic Analyses of the Microsporidian *Nosema ceranae*, an Emergent Pathogen of Honey Bees. PLoS Pathog 5:e1000466. https://doi.org/10.1371/journal.ppat.1000466

Corradi N, Haag KL, Pombert J-F, et al (2009) Draft genome sequence of the *Daphnia* pathogen *Octosporea bayeri:* insights into the gene content of a large microsporidian genome and a model for host-parasite interactions. Genome biology 10:R106–R106. https://doi.org/10.1186/gb-2009-10-10-r106

Corradi N, Pombert J-F, Farinelli L, et al (2010) The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*. Nature communications 1:77. https://doi.org/10.1038/ncomms1082

Corsaro D, Wylezich C, Venditti D, et al (2019) Filling gaps in the microsporidian tree: rDNA phylogeny of *Chytridiopsis typographi* (Microsporidia: Chytridiopsida). Parasitology Research 118:169–180. https://doi.org/10.1007/s00436-018-6130-1

Coyle CM, Weiss LM, Rhodes LV 3rd, et al (2004) Fatal myositis due to the microsporidian *Brachiola algerae*, a mosquito pathogen. N Engl J Med 351:42–47. https://doi.org/10.1056/NEJMoa032655

Cuomo CA, Desjardins CA, Bakowski MA, et al (2012) Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth. Genome Research 22:2478–2488

de Albuquerque NRM, Ebert D, Haag KL (2020) Transposable element abundance correlates with mode of transmission in microsporidian parasites. Mob DNA 11:19. https://doi.org/10.1186/s13100-020-00218-8

Dean P, Sendra KM, Williams TA, et al (2018) Transporter gene acquisition and innovation in the evolution of Microsporidia intracellular parasites. Nat Commun 9:1709. https://doi.org/10.1038/s41467-018-03923-4

Desjardins CA, Sanscrainte ND, Goldberg JM, et al (2015) Contrasting host–pathogen interactions and genome evolution in two generalist and specialist microsporidian pathogens of mosquitoes. Nature Communications 6:7121. https://doi.org/10.1038/ncomms8121

Dia N, Lavie L, Faye N, et al (2016) Subtelomere organization in the genome of the microsporidian *Encephalitozoon cuniculi*: patterns of repeated sequences and physicochemical signatures. BMC genomics 17:34–34. https://doi.org/10.1186/s12864-015-1920-7

Dia N, Lavie L, Méténier G, et al (2007) InterB multigenic family, a gene repertoire associated with subterminal chromosome regions of *Encephalitozoon cuniculi* and conserved in several human-infecting microsporidian species. Current genetics 51:171–186. https://doi.org/10.1007/s00294-006-0114-x

Didier ES, Weiss LM (2006) Microsporidiosis: current status. Current opinion in infectious diseases 19:485–492. https://doi.org/10.1097/01.qco.0000244055.46382.23

Dolgikh VV, Senderskiy IV, Pavlova OA, et al (2011) Immunolocalization of an alternative respiratory chain in *Antonospora* (*Paranosema*) *locustae* spores: Mitosomes retain their role in microsporidial energy metabolism. Eukaryotic Cell 10:588–593

Duffieux F, Peyret P, Roe BA, Vivares CP (1998) First report on the systematic sequencing of the small genome of *Encephalitozoon cuniculi* (Protozoa, Microspora): gene organization of a 4.3 kbp region on chromosome I. Microb Comp Genomics 3:1–11. https://doi.org/10.1089/omi.1.1998.3.1

Ehrenbolger K, Jespersen N, Sharma H, et al (2020) Differences in structure and hibernation mechanism highlight diversification of the microsporidian ribosome. PLoS Biol 18:e3000958. https://doi.org/10.1371/journal.pbio.3000958

Elliott TA, Gregory TR (2015) What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philosophical transactions of the Royal Society of London Series B, Biological sciences 370:20140331. https://doi.org/10.1098/rstb.2014.0331

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:238. https://doi.org/10.1186/s13059-019-1832-y

Franchet A, Niehus S, Caravello G, Ferrandon D (2019) Phosphatidic acid as a limiting host metabolite for the proliferation of the microsporidium *Tubulinosema ratisbonensis* in *Drosophila* flies. Nature Microbiology 4:645–655. https://doi.org/10.1038/s41564-018-0344-y

Freibert S-A, Goldberg AV, Hacker C, et al (2017) Evolutionary conservation and in vitro reconstitution of microsporidian iron–sulfur cluster biosynthesis. Nature Communications 8:13932. https://doi.org/10.1038/ncomms13932

Galindo LJ, Torruella G, Moreira D, et al (2018) Evolutionary Genomics of *Metchnikovella incurvata* (Metchnikovellidae): An Early Branching Microsporidium. Genome Biol Evol 10:2736–2748. https://doi.org/10.1093/gbe/evy205

Haag KL, James TY, Pombert J-F, et al (2014) Evolution of a morphological novelty occurred before genome compaction in a lineage of extreme parasites. Proc Natl Acad Sci U S A 111:15480–15485. https://doi.org/10.1073/pnas.1410442111

Haag KL, Pombert J-F, Sun Y, et al (2020) Microsporidia with Vertical Transmission Were Likely Shaped by Nonadaptive Processes. Genome Biol Evol 12:3599–3614. https://doi.org/10.1093/gbe/evz270

Heinz E, Williams TA, Nakjang S, et al (2012) The genome of the obligate intracellular parasite *Trachipleistophora hominis*: new insights into microsporidian genome dynamics and reductive evolution. PLoS Pathog 8:e1002979. https://doi.org/10.1371/journal.ppat.1002979

Hirt RP, Healy B, Vossbrinck CR, et al (1997) A mitochondrial Hsp70 orthologue in *Vairimorpha necatrix*: molecular evidence that microsporidia once contained mitochondria. Curr Biol 7:995–998. https://doi.org/10.1016/s0960-9822(06)00420-9

Hirt RP, Logsdon JMJ, Healy B, et al (1999) Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc Natl Acad Sci U S A 96:580–585. https://doi.org/10.1073/pnas.96.2.580

Huang Q, Wu ZH, Li WF, et al (2021) Genome and Evolutionary Analysis of *Nosema ceranae:* A Microsporidian Parasite of Honey Bees. Front Microbiol 12:645353. https://doi.org/10.3389/fmicb.2021.645353

Ishihara R, Hayashi Y (1968) Some properties of ribosomes from the sporoplasm of *Nosema bombycis*. Journal of Invertebrate Pathology 11:377–385. https://doi.org/10.1016/0022-2011(68)90186-9

James TY, Kauff F, Schoch CL, et al (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature 443:818–822. https://doi.org/10.1038/nature05110

James TY, Pelin A, Bonen L, et al (2013) Shared signatures of parasitism and phylogenomics unite Cryptomycota and microsporidia. Curr Biol 23:1548–1553. https://doi.org/10.1016/j.cub.2013.06.057

Katinka MD, Duprat S, Cornillot E, et al (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi.* Nature 414:450–453. https://doi.org/10.1038/35106579

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 30:3059–3066. https://doi.org/10.1093/nar/gkf436

Keeling PJ, Corradi N, Morrison HG, et al (2010) The reduced genome of the parasitic microsporidian *Enterocytozoon bieneusi* lacks genes for core carbon metabolism. Genome Biol Evol 2:304–309. https://doi.org/10.1093/gbe/evq022

Koonin EV, Fedorova ND, Jackson JD, et al (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biology 5:R7. https://doi.org/10.1186/gb-2004-5-2-r7

Ku C, Martin WF (2016) A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70 % rule. BMC Biol 14:89. https://doi.org/10.1186/s12915-016-0315-9

Lee SC, Heitman J (2017) Dynamics of parasitophorous vacuoles formed by the microsporidian pathogen *Encephalitozoon cuniculi.* Fungal genetics and biology : FG & B 107:20–23. https://doi.org/10.1016/j.fgb.2017.07.006

Leger MM, Eme L, Stairs CW, Roger AJ (2018) Demystifying Eukaryote Lateral Gene Transfer (Response to Martin 2017 DOI: 10.1002/bies.201700115). Bioessays 40:e1700242. https://doi.org/10.1002/bies.201700242

Lord JC, Vossbrinck CR, Wilson JD (2010) Occurrence of *Nosema oryzaephili* in *Cryptolestes ferrugineus* and transfer to the genus *Paranosema.* J Invertebr Pathol 105:112–115. https://doi.org/10.1016/j.jip.2010.05.005

Luallen RJ, Reinke AW, Tong L, et al (2016) Discovery of a Natural Microsporidian Pathogen with a Broad Tissue Tropism in *Caenorhabditis elegans*. PLoS Pathog 12:e1005724. https://doi.org/10.1371/journal.ppat.1005724

Lynch M, Conery JS (2003) The Origins of Genome Complexity. Science 302:1401. https://doi.org/10.1126/science.1089370

Major P, Embley TM, Williams TA (2017) Phylogenetic Diversity of NTT Nucleotide Transport Proteins in Free-Living and Parasitic Bacteria and Eukaryotes. Genome Biol Evol 9:480–487. https://doi.org/10.1093/gbe/evx015

Major P, Sendra KM, Dean P, et al (2019) A new family of cell surface located purine transporters in Microsporidia and related fungal endoparasites. Elife 8:. https://doi.org/10.7554/eLife.47037

Martin WF (2017) Too Much Eukaryote LGT. Bioessays 39:. https://doi.org/10.1002/bies.201700115

Melnikov SV, Manakongtreecheep K, Rivera KD, et al (2018) Muller’s Ratchet and Ribosome Degeneration in the Obligate Intracellular Parasites Microsporidia. Int J Mol Sci 19:. https://doi.org/10.3390/ijms19124125

Mikhailov KV, Simdyanov TG, Aleoshin VV (2017) Genomic Survey of a Hyperparasitic Microsporidian *Amphiamblys* sp. (Metchnikovellidae). Genome Biol Evol 9:454–467. https://doi.org/10.1093/gbe/evw235

Naito K, Cho E, Yang G, et al (2006) Dramatic amplification of a rice transposable element during recent domestication. Proc Natl Acad Sci USA 103:17620. https://doi.org/10.1073/pnas.0605421103

Nakjang S, Williams TA, Heinz E, et al (2013) Reduction and expansion in microsporidian genome evolution: new insights from comparative genomics. Genome Biol Evol 5:2285–2303. https://doi.org/10.1093/gbe/evt184

Nassonova ES, Bondarenko NI, Paskerova GG, et al (2021) Evolutionary relationships of *Metchnikovella dogieli* Paskerova et al., 2016 (Microsporidia: Metchnikovellidae) revealed by multigene phylogenetic analysis. Parasitol Res 120:525–534. https://doi.org/10.1007/s00436-020-06976-x

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 32:268–274. https://doi.org/10.1093/molbev/msu300

Ohta T (1992) The Nearly Neutral Theory of Molecular Evolution. Annu Rev Ecol Syst 23:263–286. https://doi.org/10.1146/annurev.es.23.110192.001403

Pan G, Xu J, Li T, et al (2013) Comparative genomics of parasitic silkworm microsporidia reveal an association between genome expansion and host adaptation. BMC Genomics 14:186. https://doi.org/10.1186/1471-2164-14-186

Parisot N, Pelin A, Gasc C, et al (2014) Microsporidian Genomes Harbor a Diverse Array of Transposable Elements that Demonstrate an Ancestry of Horizontal Exchange with Metazoans. Genome Biology and Evolution 6:2289–2300. https://doi.org/10.1093/gbe/evu178

Pelin A, Moteshareie H, Sak B, et al (2016) The genome of an *Encephalitozoon cuniculi* type III strain reveals insights into the genetic diversity and mode of reproduction of a ubiquitous vertebrate pathogen. Heredity 116:458–465. https://doi.org/10.1038/hdy.2016.4

Pelin A, Selman M, Aris-Brosou S, et al (2015) Genome analyses suggest the presence of polyploidy and recent human-driven expansions in eight global populations of the honeybee pathogen *N osema ceranae*: Genome diversity in the honeybee pathogen *Nosema ceranae*. Environ Microbiol 17:4443–4458. https://doi.org/10.1111/1462-2920.12883

Peters MJ, Suwannapong G, Pelin A, Corradi N (2019) Genetic and Genome Analyses Reveal Genetically Distinct Populations of the Bee Pathogen *Nosema ceranae* from Thailand. Microb Ecol 77:877–889. https://doi.org/10.1007/s00248-018-1268-z

Peyret P, Katinka MD, Duprat S, et al (2001) Sequence and analysis of chromosome I of the amitochondriate intracellular parasite *Encephalitozoon cuniculi* (Microspora). Genome Res 11:198–207. https://doi.org/10.1101/gr.164301

Peyretaillade E, Parisot N, Polonais V, et al (2012) Annotation of microsporidian genomes using transcriptional signals. Nature Communications 3:1137. https://doi.org/10.1038/ncomms2156

Pombert J-F, Haag KL, Beidas S, et al (2015) The *Ordospora colligata* genome: Evolution of extreme reduction in microsporidia and host-to-parasite horizontal gene transfer. mBio 6:. https://doi.org/10.1128/mBio.02400-14

Pombert J-F, Selman M, Burki F, et al (2012) Gain and loss of multiple functionally related, horizontally transferred genes in the reduced genomes of two microsporidian parasites. Proceedings of the National Academy of Sciences of the United States of America 109:12638–12643. https://doi.org/10.1073/pnas.1205020109

Pombert J-F, Xu J, Smith DR, et al (2013) Complete genome sequences from three genetically distinct strains reveal high intraspecies genetic diversity in the microsporidian *Encephalitozoon cuniculi*. Eukaryotic cell 12:503–511. https://doi.org/10.1128/EC.00312-12

Quandt CA, Beaudet D, Corsaro D, et al (2017) The genome of an intranuclear parasite, *Paramicrosporidium saccamoebae,* reveals alternative adaptations to obligate intracellular parasitism. Elife 6:. https://doi.org/10.7554/eLife.29594

Reinke AW, Balla KM, Bennett EJ, Troemel ER (2017) Identification of microsporidia host-exposed proteins reveals a repertoire of rapidly evolving proteins. Nature Communications 8:14023. https://doi.org/10.1038/ncomms14023

Roger AJ (1999) Reconstructing Early Events in Eukaryotic Evolution. Am Nat 154:S146–S163. https://doi.org/10.1086/303290

Roy B, Granas D, Bragg F, et al (2020) Autoregulation of yeast ribosomal proteins discovered by efficient search for feedback regulation. Communications Biology 3:761. https://doi.org/10.1038/s42003-020-01494-z

Selman M, Pombert J-F, Solter L, et al (2011) Acquisition of an animal gene by microsporidian intracellular parasites. Current biology : CB 21:R576-7

Selman M, Sak B, Kváč M, et al (2013) Extremely reduced levels of heterozygosity in the vertebrate pathogen *Encephalitozoon cuniculi*. Eukaryotic cell 12:496–502. https://doi.org/10.1128/EC.00307-12

Sokolova YY, Issi IV, Morzhina EV, et al (2005) Ultrastructural analysis supports transferring *Nosema whitei* Weiser 1953 to the genus *Paranosema* and creation a new combination, *Paranosema whitei.* J Invertebr Pathol 90:122–126. https://doi.org/10.1016/j.jip.2005.06.009

Thitamadee S, Prachumwat A, Srisala J, et al (2016) Review of current disease threats for cultivated penaeid shrimp in Asia. Aquaculture 452:69–87. https://doi.org/10.1016/j.aquaculture.2015.10.028

Tiessen A, Pérez-Rodríguez P, Delaye-Arredondo LJ (2012) Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. BMC Research Notes 5:85. https://doi.org/10.1186/1756-0500-5-85

Timofeev S, Tokarev Y, Dolgikh V (2020) Energy metabolism and its evolution in Microsporidia and allied taxa. Parasitology Research 119:1433–1441. https://doi.org/10.1007/s00436-020-06657-9

Tokarev YS, Huang W-F, Solter LF, et al (2020) A formal redefinition of the genera *Nosema* and *Vairimorpha* (Microsporidia: Nosematidae) and reassignment of species based on molecular phylogenetics. J Invertebr Pathol 169:107279. https://doi.org/10.1016/j.jip.2019.107279

Troemel ER, Becnel JJ (2015) Genome analysis and polar tube firing dynamics of mosquito-infecting microsporidia. Fungal Genet Biol 83:41–44. https://doi.org/10.1016/j.fgb.2015.08.007

Tsaousis AD, Kunji ERS, Goldberg AV, et al (2008) A novel route for ATP acquisition by the remnant mitochondria of *Encephalitozoon cuniculi*. Nature 453:553–556. https://doi.org/10.1038/nature06903

Van Etten J, Bhattacharya D (2020) Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? Trends Genet 36:915–925. https://doi.org/10.1016/j.tig.2020.08.006

Watson AK, Williams TA, Williams BAP, et al (2015) Transcriptomic profiling of host-parasite interactions in the microsporidian *Trachipleistophora hominis*. BMC Genomics 16:. https://doi.org/10.1186/s12864-015-1989-z

Whelan TA, Lee NT, Lee RCH, Fast NM (2019) Microsporidian Introns Retained against a Background of Genome Reduction: Characterization of an Unusual Set of Introns. Genome biology and evolution 11:263–269. https://doi.org/10.1093/gbe/evy260

Williams BAP, Hirt RP, Lucocq JM, Embley TM (2002) A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*. Nature 418:. https://doi.org/10.1038/nature00949

Williams TA, Nakjang S, Campbell SE, et al (2016) A Recent Whole-Genome Duplication Divides Populations of a Globally Distributed Microsporidian. Molecular Biology and Evolution 33:. https://doi.org/10.1093/molbev/msw083

Wiredu Boakye D, Jaroenlak P, Prachumwat A, et al (2017) Decay of the glycolytic pathway and adaptation to intranuclear parasitism within Enterocytozoonidae microsporidia. Environmental Microbiology 19:. https://doi.org/10.1111/1462-2920.13734

Xiao L, Li L, Visvesvara GS, et al (2001) Genotyping *Encephalitozoon cuniculi* by multilocus analyses of genes with repetitive sequences. J Clin Microbiol 39:2248–2253. https://doi.org/10.1128/JCM.39.6.2248-2253.2001

Xu J, He Q, Ma Z, et al (2016) The Genome of *Nosema* sp. Isolate YNPr: A Comparative Analysis of Genome Evolution within the *Nosema/Vairimorpha* Clade. PloS one 11:e0162336–e0162336. https://doi.org/10.1371/journal.pone.0162336

Zhang J (2000) Protein-length distributions for the three domains of life. Trends Genet 16:107–109. https://doi.org/10.1016/s0168-9525(99)01922-8

Figure Legends:

Figure 1: Accumulation of microsporidian genomes assemblies over time in NCBI. The top line shows accumulation of numbers of assemblies whilst the lower line indicates the number of new unique species assemblies. X axis shows years and Y axis, numbers of available genomes.

Figure 2: Phylogenetic relationships between microsporidian species with sequenced genomes with to the right-hand side indication of genome/assembly size with proportion taken up by coding sequences in black (Note that for some species the value for some figures the value was taken from published estimates which may not include gene duplicates). ♦ indicates putative loss of glycolysis, ✶ indicates putative loss of introns and a functional spliceosome ✕ indicates putative losses of loss of the RNA interference pathway. (Available proteomes for sequenced genomes at NCBI were downloaded on the 18th May 2021. For *A/P. locustae* and *Me. incurvata* and *Spraguea* NB2 the nucleotide assembly was downloaded and EMBOSS getorf (min 100) used to predict proteins. OrthoFinder (default settings) was used to create orthogroups (Emms and Kelly 2019). Any orthogroups with representation for at least 20 taxa were selected and aligned using MAFFT and trimmed using trimAl (Katoh et al. 2002; Capella-Gutiérrez et al. 2009). Initial trees (IQ-TREE) were inspected and close protein orthologs that were the result of duplication within a single species were removed to leave a single copy (Nguyen et al. 2015). Alignments were concatenated and IQ-TREE was run on the partitioned file with best model indicated for each aligned protein and 1000 UltraFast bootstrap replicates. Values for bootstraps were 100% unless indicated.).

Figure 3: Putative *de novo* gene area in *Spraguea* spp.: Shown is the same area of the genome of two populations of the same pathogen. In the East Atlantic genome, it is possible to see and area with an open reading frame that is transcribed and putatively translated into a population specific protein. In the West Atlantic genome, the same area does not contain an open reading frame. Differences at the DNA level between these two populations are indicated by black boxes.

Figure 4: Polar tube 2 region of the various genomes to illustrate variation in intergenic spacing between species. Shown is the PTP2 gene shaded in grey which is chosen as a reference point. Surrounding genes are shown in white. Homologous genes are indicated by light grey shading linking genes between species. Relationships between species shown to the left hand side are extracted from the phylogeny in Figure 2. The hatched box in the second genomic area for *Nosema bombycis* indicates a possible pseudogene. It is also possible to see the synteny within *Encephalitozoon* and *Nosema* species but this synteny breaks down with increasing phylogenetic distance which may reflect increased time since divergence but also a higher chance of fixation of recombinations as genome density decreases and the chance of disrupting an existing open reading frame decreases.

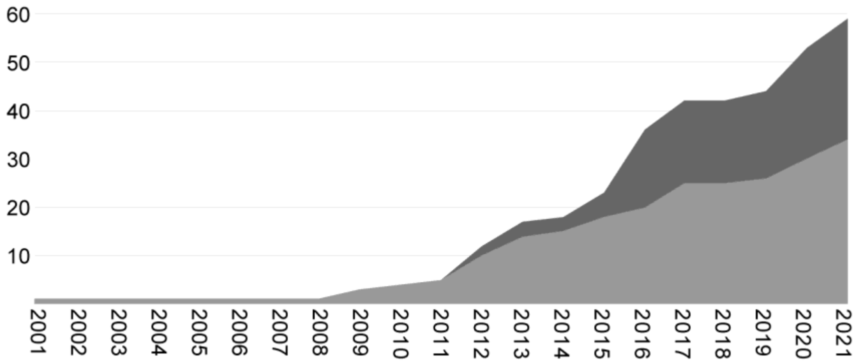

60
50
40
30
20
10
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

Figure 1

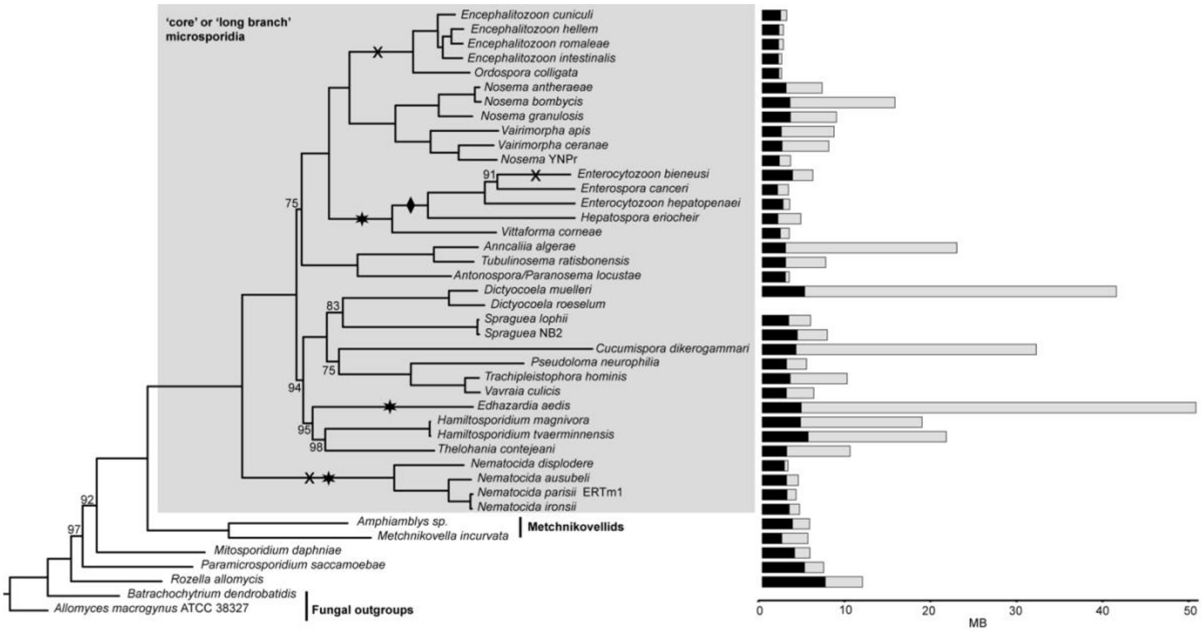
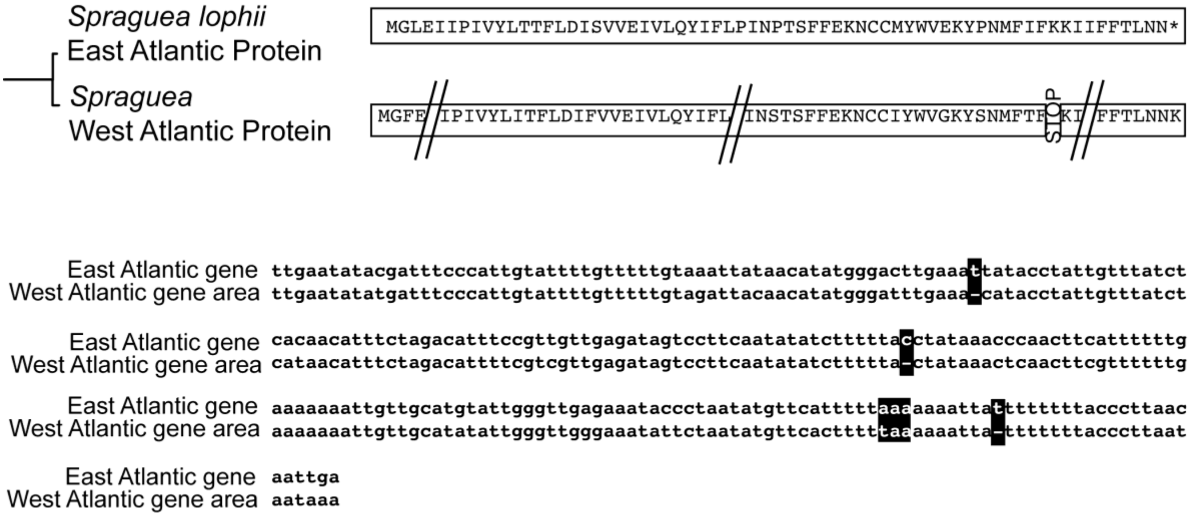

'core' or 'long branch' microsporidia
Encephalitozoon cuniculi
Encephalitozoon hellem
Encephalitozoon romaleae
Encephalitozoon intestinalis
Ordospora colligata
Nosema antheraeae
Nosema bombycis
Nosema granulosis
Vairimorpha apis
Vairimorpha ceranae
Nosema YNPr
91
Enterocytozoon bieneusi
Enterospora canceri
Enterocytozoon hepatopenaei
Hepatospora eriocheir
75
Vittaforma corneae
Anncaliia algerae
Tubulinosema ratisbonensis
Antonospora/Paranosema locustae
Dictyocoela muelleri
83
Dictyocoela roeselum
Spraguea lophii
Spraguea NB2
Cucumispora dikerogammari
75
Pseudoloma neurophilia
Trachipleistophora hominis
94
Vavraia culicis
Edhazardia aedis
95
Hamiltosporidium magnivora
Hamiltosporidium tvaerminnensis
98
Thelohania contejeani
Nematocida displodere
Nematocida ausubeli
Nematocida parisii ERTm1
92
Nematocida ironsii
Amphiamblys sp.
97
Metchnikovella incurvata
Metchnikovellids
Mitosporidium daphniae
Paramicrosporidium saccamoebae
Rozella allomycis
Batrachochytrium dendrobatidis
Allomyces macrogynus ATCC 38327
Fungal outgroups
0
10
20
30
40
50
MB

Figure 2


Spraguea lophii
East Atlantic Protein
Spraguea
West Atlantic Protein
MGLEIIPIVYLTTFLDISVVEIVLQYIFLPINPTSFFEKNCCMYWVEKYPNMFIFKKIIFFTLNN*
MGFE IPIVYLITFLDIFVVEIVLQYIFL INSTSFFEKNCCIYWVGKYSNMFTF STOP KI FFTLNNK

```
     East Atlantic gene ttgaatatacgatttcccattgtattttgtttttgtaaattataacatatgggacttgaaattatacctattgtttatct
West Atlantic gene area ttgaatatatgatttcccattgtattttgtttttgtagattacaacatatgggatttgaaa-catacctattgtttatct

     East Atlantic gene cacaacatttctagacatttccgttgttgagatagtccttcaatatatctttttacctataaacccaacttcatttttg
West Atlantic gene area cataacatttctagacattttcgtcgttgagatagtccttcaatatatcttttta-ctataaactcaacttcgtttttg

     East Atlantic gene aaaaaaattgttgcatgtattgggttgagaaataccctaatatgttcattttaaaaaaattatttttttacccttaac
West Atlantic gene area aaaaaaattgttgcatatattgggttgggaaatattctaatatgttcacttttaaaaaaatta-tttttttacccttaat

     East Atlantic gene aattga
West Atlantic gene area aataaa
```

Figure 3

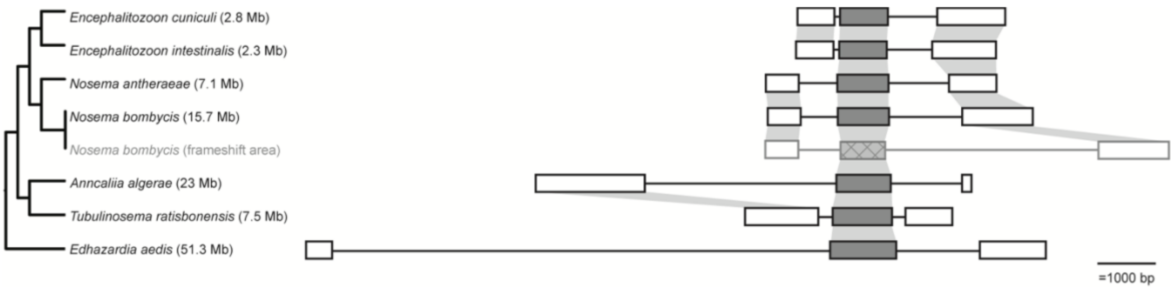

Encephalitozoon cuniculi (2.8 Mb)
Encephalitozoon intestinalis (2.3 Mb)
Nosema antheraeae (7.1 Mb)
Nosema bombycis (15.7 Mb)
Nosema bombycis (frameshift area)
Anncaliia algerae (23 Mb)
Tubulinosema ratisbonensis (7.5 Mb)
Edhazardia aedis (51.3 Mb)
=1000 bp

Figure 4