

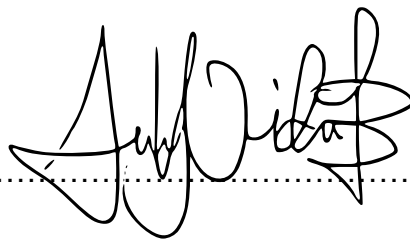
Life in the nucleus, the genomic basis of energy exploitation by intranuclear
microsporidia

Submitted by Dominic Wiredu Boakye to the University of Exeter
As a thesis for the degree of
Doctor of Philosophy in Biological Sciences
In September 2016

This thesis is available for library use on the understanding that it is copyright
Material and that no quotation from the thesis may be published without proper
acknowledgement.

I certify that all material in this thesis which is not my own work has been
Identified and that no material has previously been submitted and approved for
The award of a degree by this or any other University.

Signature:

A handwritten signature in black ink, appearing to read 'Dominic Wiredu Boakye', written over a horizontal dotted line.

Abstract

The Microsporidia are obligate intracellular parasites that have jettisoned oxidation phosphorylative capabilities during their early evolutionary history and so rely on ATP import from their host and glycolysis for their energy needs. Some species form tight associations with the host's mitochondria and this is thought to facilitate ATP sequestration by the developing intracellular microsporidian. The human parasite, *Enterocytozoon bieneusi* has however lost glycolytic capabilities and may rely entirely on ATP import from its host for energy. *E. bieneusi* belongs to the Enterocytozoonidae microsporidian family and recent rDNA-based phylogenetic studies have suggested it has close evolutionary ties with *Enterospora canceri*, a crab-infecting intranuclear parasite. Such a close evolutionary relationship implied that glycolysis might also be absent in the intranuclear parasite raising questions as to how this parasite obtains energy from its unusual niche that is physically walled off from the host mitochondria, the main source of ATP in the host cell.

In this study, draft genomes of four species of the Enterocytozoonidae namely, *Ent. canceri*, *E. hepatopenaei*, *Hepatospora eriocheir* and *Hepatospora eriocheir canceri* and one non-Enterocytozoonidae species, *Thelohania* sp. were assembled and annotated (The genome assembly of *Hepatospora eriocheir* was provided by Dr. Bryony Williams). Phylogenomics performed with this and publicly available genomic data confirmed the close evolutionary ties between *Ent. canceri* and *E. bieneusi*. Comparative genomic analyses also revealed that glycolysis is indeed lost in all members of the Enterocytozoonidae family sequenced in this study, hinting to the relaxation of evolutionary pressures to maintain this pathway at the base of this microsporidian family. Despite this absence, the hexokinase gene was retained in all aglycolytic genomes analysed, and that of *Ent. canceri* was fused to a PTPA gene. Functional assays and yeast complementation assays suggest that this chimera is able to recognise glucose as a substrate but the heterologously expressed homolog of *H. eriocheir* cannot. Finally, phylogenomics have been used here to demonstrate that despite the morphological differences between three *Hepatospora*-like organisms parasitizing different crab hosts, they are the same species. This finding adds more weight to current evidence suggesting that morphology is not an ideal marker for taxonomical classification in the Microsporidia.

Acknowledgements

I would like to thank my supervisors, Dr. Bryony Williams and Prof. Grant Stentiford for allowing me to be part of this amazing project and mentoring me for the past four years.

I would also like to thank Dr. Thomas Richards and his group with whom I was privileged to work, albeit for a short period for the bioinformatics assistance they have offered me through out my Ph. D. In addition, thanks go to Tom Williams (University of Bristol) who guided me to construct my first phylogenetic tree. I would also like to thank Konrad Paszkiewicz and Paul O'Neil (University of Exeter sequencing service) for their assistance with regards to writing Perl and Unix scripts and performing *de novo* genome assemblies.

Thank you to Dr. Kelly Bateman (CEFAS) for teaching me all I know about crab dissections and to Pattana Jaroenlak (Mahidol University, Thailand) for the provision of EHP spores, a vital part of this project. Dr. Sophie Shaw, Dr. Lauren Ames, Dr. Graham Thomas, Dr. Jane Usher, Josephine Paris, Thomas Jenkins, Guy Freeman, Dr. Guy Leonard, Dr. Adam Monier, Dr. Dogra Yuktee, Yogesh Chaudhari and David Milner, thank you for proofreading my thesis, offering me temporary accommodation when I needed it, making me cake, ordering synthetic genes, helping me with R scripts and for being such incredible friends.

Then there are my parents Jones Wiredu Boakye and Esther Sarfowa who have been very supportive throughout these four years, Thank You. Finally, a big Thank You to my beautiful wife, Monica Lamici Ayine for putting up with my stressing at home and for being a constant source of motivation.

Table of Contents

Abstract	2
Acknowledgements	3
Table of figures	15
List of Tables	18
Abbreviations	20
Chapter 1 General Introduction	23
1.1 The Microsporidia: structure and development	23
1.1.1 The spore	23
1.1.2 The meront	26
1.1.3 The sporont	26
1.2 History of the phylogenetic placement of the Microsporidia	28
1.2.1 Coining of the name Microsporidia and addition to the Schizomycetes	28
1.2.2 The Archezoa hypothesis.....	28
1.2.3 The mitosome and the end of the Archezoa hypothesis	29
1.2.4 The phylogenetic link between the Microsporidia and Fungi	30
1.3 Systematics within the Microsporidia	33
1.3.1 Creation of the order Microsporides	33
1.3.2 The use of phenotypic traits for taxonomic ranking in the	
Microsporidia	33
1.3.3 The use of ultrastructural features for taxonomic ranking in the	
Microsporidia	34
1.3.4 The phylum Microsporidia	34
1.3.5 Ribosomal DNA and systematics of the Microsporidia	35
1.4 The reduced microsporidian genome and intracellular living	38
1.4.1 Obligate intracellular parasitism: A consequence of an ancient	
facultative relationship.....	38
1.4.2 The evolutionary trajectory of intracellular lineages is dictated by cell	
cycle and host environment	39
1.4.3 Factors contributing to genome shrinkage in intracellular parasites .	39

1.4.3.1 Cell size	40
1.4.3.2 Metabolic Economy	41
1.4.3.3 Genome maintenance requirements	41
1.4.3.4 Effective population size decline and genetic drift:	43
1.4.3.5 Loss of genetic redundancy.....	43
1.4.3.6 Other contributors to the reduced genome in the Microsporidia.	44
1.5 The mitochondria and energy conservation	44
1.6 Absence of the oxidative phosphorylation pathway in the Microsporidia	46
1.7 The role of microsporidian-host-mitochondria association in ATP acquisition	46
1.8 The Enterocytozoonidae family	47
1.8.1 Cytoplasmic-infecting species of the Enterocytozoonidae	49
1.8.1.1 <i>Desmozoon lepeophtherii</i> (= <i>Paranucleospora theridion</i>)	50
1.8.1.2 <i>Hepatospora</i> spp.	51
1.8.1.3 <i>Enterocytozoon hepatopenaei</i>	53
1.8.1.4 <i>Enterocytozoon bieneusi</i>	54
1.8.2 Intranuclear species of the Enterocytozoonidae	56
1.8.2.1 <i>Nucleospora</i> spp.....	57
1.8.2.2 <i>Enterospora nucleophila</i>	58
1.8.2.3 <i>Enterospora canceri</i>	60
1.9 The nucleus as a niche	62
1.10 Importance of studying the Enterocytozoonidae family	63
1.10.1 Potential for zoonotic transmission in humans	63
1.10.2 Diseases of commercially important fisheries	64
1.10.3 Potential biological control agents.....	65
1.10.4 Models for studying eukaryotic extreme genome minimalism.....	66
1.11 Overall aims and objectives of study	67
1.11.1 Understand the extent of metabolic loss in the Enterocytozoonidae	67
1.11.2 Identify signatures of intranuclear living	67
1.11.3 Phylogenetic assessment of the <i>Hepatospora</i> genus	68

Chapter 2 Comparative genomics of the Enterocytozoonidae reveals extreme loss in metabolic capacity **70**

2.1 Introduction	70
------------------	----

2.1.1 Absence of oxidation phosphorylative pathways in the Microsporidia	70
2.1.2 Absence of core metabolic genes in the genome of <i>Enterocytozoon bieneusi</i>	70
2.1.3 Host nucleus associations in microsporidian infections	71
2.1.4 Assembling Next Generation Sequencing (NGS) data	72
2.2 Main aims of study	75
2.3 Methods	76
2.3.1 Sampling of edible crabs	76
2.3.2 Identification of edible crab infected tissues	76
2.3.3 <i>Enterospora canceri</i> and <i>Hepatospora eriocheir canceri</i> spore isolation from the edible crab	76
2.3.4 <i>Thelohania</i> sp. spore isolation from European crayfish claws	76
2.3.5 Acquiring spores of <i>Enterocytozoon hepatopenaei</i>	77
2.3.6 Genomic DNA extraction for sequencing	77
2.3.7 Assembling microsporidian genomes	77
2.3.7.1 Preliminary assemblies	77
2.3.7.2 Assembly optimization by Illumina read GC content filtering	77
2.3.7.2.1 SPADES Assembly	78
2.3.7.2.2 A5 MISEQ Assembly:	78
2.3.7.2.3 VELVET Assembly	79
2.3.7.2.4 RAY Assembly	79
2.3.7.2.5 Selecting the best preliminary assembly	79
2.3.7.3 Reassembling GC-filtered preliminary assemblies	81
2.3.7.4 Comparing predicted protein sets in the genomes of <i>Hepatospora eriocheir</i> and <i>Hepatospora eriocheir canceri</i>	81
2.3.8 Assembled genome of <i>Hepatospora eriocheir</i>	82
2.3.9 MAKER genome annotation	82
2.3.9.1 Masking repetitive regions	82
2.3.9.2 Ab-initio gene calling	83
2.3.9.3 BLAST-based gene calling	83
2.3.10 Formatting annotated genomes for GENBANK submission.	83
2.3.10.1 Assigning names to MAKER-predicted genes	83
2.3.10.2 Formatting MAKER-predicted proteins for SEQUIN submission	84

2.3.10.3	Generation of a preliminary feature table file with SEQUIN	85
2.3.10.4	Generation of final sequin file for GENBANK submission	88
2.3.10.5	Identification of ORFs missed by MAKER	89
2.3.11	Identification of gene families	89
2.3.12	Phylogenomic assessment of the microsporidian phylum	90
2.3.12.1	Concatenated protein alignment	90
2.3.12.2	Maximum likelihood analysis on 21-protein concatenated alignment	90
2.3.12.3	Bayesian inference analysis on 21-protein concatenated alignment	90
2.3.13	Comparative genomic analysis	91
2.3.13.1	Genomes used in this study	91
2.3.13.2	Identifying core microsporidian proteins in the newly sequenced genomes	91
2.3.13.3	Identifying putative transcription factor binding domains	92
2.3.13.4	Scanning genomes for transposable elements and tRNAs	92
2.3.13.5	Assessment of synonymous codon usage and codon usage bias within sequenced Enterocytozoonidae genomes	93
2.3.13.6	Identification of transporter proteins	93
2.3.13.7	Identification of secreted proteins	94
2.4	Results	94
2.4.1	Assembly and optimization of <i>Hepatospora eriocheir canceri</i> genome	94
2.4.1.1	Preliminary assembly	94
2.4.1.2	Optimizing assembly by GC-content Illumina read filtering	95
2.4.1.2.1	SPADES assembly	95
2.4.1.2.2	A5-MISEQ assembly	95
2.4.1.2.3	VELVET assembly	96
2.4.1.2.4	RAY assembly	96
2.4.1.2.5	Selecting best assembly program for the genome of <i>Hepatospora eriocheir canceri</i>	97
2.4.1.2.6	Selecting the best GC content cut-off for the genome of <i>Hepatospora eriocheir canceri</i>	98
2.4.1.2.7	Further assembly optimization by aligning genomes of the two <i>Hepatospora</i> spp. to each other	99

2.4.2 Assembly and optimization of <i>Enterocytozoon hepatopenaei</i> genome	100
2.4.2.1 Preliminary assembly	100
2.4.2.2 Optimizing assembly by Illumina read GC-content filtering	101
2.4.2.2.1 SPADES assembly	101
2.4.2.2.2 A5-MISEQ assembly	102
2.4.2.3 Selecting best assembly program for the genome of <i>Enterocytozoon hepatopenaei</i>	103
2.4.2.4 Optimizing the assembly of <i>Enterocytozoon hepatopenaei</i> by assessing read coverage	104
2.4.3 Preliminary assembly of the <i>Enterospora canceri</i> genome	105
2.4.4 Assembly optimization of the genome of <i>Enterospora canceri</i>	108
2.4.4.1 Establishing consensus assembly between RUN1 and RUN2	109
2.4.4.1.1 Filtering aligned RUN1 assembly by k-mer coverage, BLAST and contig length	109
2.4.4.1.2 Reassembling RUN1 with reads that mapped onto filtered contigs	110
2.4.4.1.3 Reassembling RUN2 with reads that mapped onto filtered contigs	111
2.4.4.1.4 De novo assembly of genome of <i>Enterospora canceri</i> by combining reads from RUN1 and RUN2	111
2.4.4.1.5 Filtering RUN3 Assembly by k-mer coverage	112
2.4.4.1.6 Assembling the genome of <i>Enterospora canceri</i> with reads from RUN1, 2 and 3	114
2.4.4.2 The final <i>Enterospora canceri</i> assembly	115
2.4.5 Predicted open reading frames (ORFs) in the sequenced genomes	115
2.4.6 Conserved motifs upstream of start codons	116
2.4.7 tRNAs found in the sequenced genomes and frequencies of their corresponding amino acids	116
2.4.8 GC content and synonymous codon usage bias in the <i>Enterocytozoonidae</i>	117
2.4.9 Identifying transposable elements and repetitive DNA sequences	120
2.4.9.1 DFAM predictions	120
2.4.9.2 REPEATFINDER predictions	120

2.4.10 Phylogenomics of the Microsporidia	121
2.4.11 Mapping of metabolic functions onto the microsporidian phylum..	121
2.4.12 Comparing the plasma membrane transporter repertoire of nuclear and cytoplasm-infecting microsporidians	123
2.4.13 Secreted proteins in the Microsporidia	125
2.5 Discussion	130
2.5.1 Variability in assembly quality is due to different heuristic approaches used by assembly programs to assess errors, inconsistency and ambiguity.	130
2.5.2 Origin of contamination in Illumina data	131
2.5.3 Low quality base pairs in raw reads of <i>Enterospora canceri</i> is as a result of poor library preparation	132
2.5.4 Decontamination protocols depend on target and contaminating genome properties	134
2.5.4.1 <i>Hepatospora eriocheir canceri</i> : A case of low target and high contaminant genomic GC content	134
2.5.4.2 <i>Enterocytozoon hepatopenaei</i> : A case of low target and contaminant genomic GC content	134
2.5.4.3 <i>Enterospora canceri</i> : a case of high GC content in target and contaminant genome	135
2.5.5 Phylogenomics of the Enterocytozoonidae	137
2.5.6 Enterocytozoonidae and transposable elements	138
2.5.6.1.1 Gypsy and Tc1 retrotransposons: Ubiquitous transposable elements in the Microsporidia	139
2.5.7 Absence of complementary tRNAs for some codons in microsporidian ORFeomes	140
2.5.8 The GGGTAAAA motif: A putative transcription binding site of the Enterocytozoonidae	141
2.5.9 Extreme reduction in metabolic capacity within the Enterocytozoonidae	142
2.5.10 Partial conservation of deoxyribonucleotide metabolism within the Microsporidia	142
2.5.11 Assessment of plasma membrane transporters to uncover a signature of intranuclear parasitism	144

2.5.12 The predicted secretomes of <i>Hepatospora</i> spp., <i>Enterocytozoon bieneusi</i> , <i>Enterospora canceri</i> and <i>Enterocytozoon hepatopenaei</i>	145
2.6 Conclusion	148
Chapter 3 The phylum Microsporidia and loss of glycolytic enzymes	149
3.1 Introduction	149
3.1.1 Universality of glycolysis and plasticity within this pathway	149
3.1.2 How the Microsporidia acquire ATP.....	151
3.1.2.1 Glycolysis	151
3.1.2.2 Horizontally acquired ATP/ADP translocases	152
3.1.3 The microsporidian hexokinase	153
3.2 Main aims of study	155
3.3 Methods	156
3.3.1 Bioinformatics.....	156
3.3.1.1 Hexokinase phylogeny	156
3.3.1.2 PTPA phylogeny	157
3.3.1.3 Mapping hexokinase active sites.....	157
3.3.1.4 Genome-wide analysis to identify chimeric proteins within the Microsporidia	157
3.3.1.5 Assessment of gene order.....	157
3.3.2 Cloning: Organisms, strains and plasmids.....	158
3.3.2.1 Bacterial strains	158
3.3.2.2 Yeast strains.....	158
3.3.3 Primers.....	158
3.3.4 Plasmids.....	158
3.3.5 Media and solutions	159
3.3.5.1 Bacterial media solutions.....	159
3.3.5.2 Yeast media.....	159
3.3.5.3 DNA electrophoresis.....	160
3.3.5.4 Solutions for protein work.....	160
3.3.6 Molecular techniques	161
3.3.6.1 Addition of attB sites onto amplified microsporidian hexokinase	161
3.3.6.2 Resolution of DNA fragments by gel electrophoresis	162
3.3.6.3 Purification of PCR products from agarose gels.....	162

3.3.6.4 BP clonase reaction.....	162
3.3.6.5 LR clonase reaction.....	162
3.3.6.6 Chemical <i>Escherichia coli</i> transformation.....	162
3.3.6.7 Isolation of plasmid DNA from <i>Escherichia coli</i> cells.....	163
3.3.6.8 Cloning poly-HIS tagged microsporidian hexokinases into PYES2 Saccharomyces cerevisiae vectors	163
3.3.6.9 Cloning microsporidian hexokinases into pAG426GPD-EGFP Saccharomyces cerevisiae vectors	164
3.3.6.10 Lithium Acetate-mediated transformation of <i>Saccharomyces cerevisiae</i>	164
3.3.6.11 Expression of microsporidian hexokinases in <i>Escherichia coli</i>	164
3.3.6.12 Expression of microsporidian hexokinases in <i>Saccharomyces cerevisiae</i>	165
3.3.6.13 Recombinant protein HIS-tag purification.....	167
3.3.6.14 Western blotting.....	167
3.3.6.15 Construction of an empty pEXP17 clone for use in BIOLOG analysis.....	167
3.3.6.16 BIOLOG Analysis	168
3.3.6.17 Hexokinase activity assay	168
3.3.6.18 Processing of protein samples for Matrix-assisted laser Desorption/Ionization Mass Spectrometry (MALDI-MS) analysis.....	169
3.4 Results	171
3.4.1 Loss of glycolysis in the Enterocytozoonidae.....	171
3.4.2 Phylogenetic assessment of microsporidian hexokinases	174
3.4.3 Phylogenetic assessment of microsporidian PTPA proteins.....	177
3.4.4 Evolution of the chimeric hexokinase in <i>Enterospora canceri</i>	178
3.4.5 Genome wide analysis of chimeric genes in the Microsporidia.....	181
3.4.6 Phylogenetic assessment of microsporidian ATP/ADP translocases	181
3.4.7 Characterization of microsporidian hexokinases.....	183
3.4.7.1 Yeast functional complementation assay	183
3.4.7.2 MALDI-MS analysis of excised 50 and 25 kDa protein bands..	184
3.4.7.3 Heterologous expression of microsporidian hexokinase in <i>Escherichia coli</i>	184

3.4.7.4 Hexokinase functional assay	185
3.4.7.5 BIOLOG phenotypic microarray	187
3.5 Discussion	189
3.5.1 Absence of glycolysis within the Enterocytozoonidae	189
3.5.2 Absence of glycolytic genes in sequenced genomes is not due to incomplete genome sequencing	189
3.5.3 Explaining events that led to the decay of glycolysis in the Enterocytozoonidae	190
3.5.3.1 Horizontal transfer of ATP/ADP translocases in the Microsporidia	190
3.5.3.2 Loss of cytosolic NAD ⁺ pool replenishing pathways	191
3.5.3.2.1 AOX enzymes	191
3.5.3.2.2 Loss of the mevalonate biosynthesis pathway	191
3.5.3.2.3 Loss of NUDIX enzymes	192
3.5.3.2.4 Loss of NAD ⁺ transport systems	193
3.5.3.3 Summary of events that may have led to loss of glycolysis in the Enterocytozoonidae	193
3.5.4 Microsporidian hexokinases	194
3.5.5 Functional characterization of hexokinase from <i>Enterospora canceri</i> and <i>Hepatospora eriocheir</i>	197
3.6 Conclusion	201

Chapter 4 *Hepatospora*-An example of plasticity in microsporidian

morphology and karyotype	202
4.1 Introduction	202
4.2 Main aims of study	207
4.3 Methods	208
4.3.1 Sampling of edible crabs	208
4.3.2 Identification of infected tissues of the edible crab	208
4.3.3 Spore extraction from infected tissues isolated from the edible crab	208
4.3.4 Genomic DNA extraction	208
4.3.5 Illumina read assembly for the edible crab parasite	208
4.3.6 Marker genes used in this study	208

4.3.7 Identification of six marker genes from assembled genomes of <i>Hepatospora eriocheir</i> and edible crab parasite	208
4.3.8 Primer design, PCR and sequencing of six marker genes from the pea crab parasite	209
4.3.8.1 Creation of a six-gene concatenated phylogenetic tree	209
4.3.8.1.1 Maximum likelihood analysis	209
4.3.8.1.2 Bayesian inference analysis on six-gene concatenated alignment.	210
4.4 Results	211
4.4.1 Creation of a six-gene concatenated phylogenetic tree	211
4.4.1.1 Individual gene trees	211
4.4.1.2 Six-gene concatenated tree.....	214
4.5 Discussion	215
4.5.1 Taxonomic names for the edible and pea crab parasites	215
4.5.2 Importance of continuous disease profile surveillance for UK fisheries	216
4.5.3 Origin of <i>Hepatospora eriocheir</i>	216
4.5.4 Limitations of study and future outlook.....	217
4.6 Conclusion	220
Chapter 5 Summary and future perspectives	221
5.1 New genomic data for four Enterocytozoonidae species: Impact on aquaculture and global food security	221
5.2 Systematics of the Enterocytozoonidae	223
5.3 Loss of glycolysis: a common trait within the Enterocytozoonidae	224
5.4 Intranuclear living may have led to gene dosage increase of plasma membrane transporters	224
Appendix 1: Assessing the taxonomic profile of 381-core-eukaryotic proteins present in the assembled genome of <i>Enterocytozoon hepatopenaei</i>	226
Appendix 2: Plasmids used in this study.....	233
Appendix 3: List of oligonucleotides used in this study	236
Appendix 4: Bash scripts and partition files for 21-protein phylogenetic analysis	238

Appendix 5: Bash scripts for plasma membrane transporter prediction	241
Appendix 6: Secretome predictions for the Microsporidia	247
Appendix 7: Secretome prediction bash scripts	249
Appendix 8: Matrix-assisted laser desorption/Ionization Mass Spectrometry (MALDI-MS) analysis	255
Appendix 9: Transposable element predictions	260
Appendix 10: Plasma membrane transporter predictions	279
Appendix 11: Taxonomic profile of <i>Enterospora canceri</i>'s RUN1 assembly	279
Appendix 12: Bateman <i>et al.</i>, 2016 Paper published in the Parasitology Journal.....	281
Appendix 13: Wiredu-Boakye <i>et al.</i>, 2016 paper submitted to PLoS Pathogens	293
Appendix 14: Phylogenomic analyses of 23 microsporidian species	339
Bibliography.....	340

Table of figures

Figure 1.1: Polymorphic nature of the microsporidian spore wall.....	24
Figure 1.2: 3D cartoon of a microsporidian spore: Summarizing the different stages (1-19) involved in polar filament extrusion	27
Figure 1.3: Schematic representation of the current view of the phylogenetic position of the Microsporidia.....	32
Figure 1.4: rDNA-based phylogeny of 92 sampled microsporidian species demonstrates that microsporidian lineages are better united by habitat	37
Figure 1.5: Relationship between genome size and relative spore size for selected microsporidian species	40
Figure 1.6: Mapping GC content and host cell longevity on genome size	42
Figure 2.1: Comparing the Assembly sizes and N50 values for the genome of <i>Hepatospora eriocheir canceri</i> produced by four eukaryotic genome assemblers at increasing Illumina read GC content cut-offs	98
Figure 2.2: Comparing completeness between genome assemblies of <i>Hepatospora eriocheir canceri</i> performed with Illumina reads filtered at various GC content cut-offs	99
Figure 2.3: Comparing genomic DNA assemblies of the two <i>Hepatospora</i> species	100
Figure 2.4: Estimating completeness of assembled genome of <i>Enterocytozoon hepatopenaei</i> at different GC content cut-offs.....	104
Figure 2.5 : Optimizing assembly of <i>Enterocytozoon hepatopenaei</i>	105
Figure 2.6: Coverage distribution of <i>Enterosporea canceri</i> 's Illumina reads on its SPADES-assembled genome from RUN1 displaying a unimodal distribution.	107
Figure 2.7: Estimating completeness of assembled genome of <i>Enterosporea canceri</i> at different GC content cut-offs.....	108
Figure 2.8: Establishing filtering parameters (dotted lines) for the genome of <i>Enterosporea canceri</i>	110
Figure 2.9: Assessing levels of contamination in genome of <i>Enterosporea canceri</i> reassembled from RUN1 and RUN2 Illumina reads.....	112

Figure 2.10: Establishing filtering parameters for the RUN3 assembly of <i>Enterospora canceri</i> 's genome.....	114
Figure 2.11: Comparing putative regulatory motifs upstream of microsporidian ORFs.	116
Figure 2.12: tRNAs in the Enterocytozoonidae	117
Figure 2.13: Frequency of each amino acid in predicted ORFs of the Enterocytozoonidae and <i>Vittaforma</i>	117
Figure 2.14: Visualization of synonymous codon usage bias (SCUO) against GC distribution for ORFs in the genomes of the Enterocytozoonidae.....	119
Figure 2.15: Comparing codon usage frequencies	120
Figure 2.16: Cladogram of 23 microsporidian species	121
Figure 2.17: Metabolic profiling of microsporidian genomes	122
Figure 2.18: Number of predicted plasma membrane transporters belonging to various protein families identified for seven microsporidian genomes	124
Figure 2.19: Comparative assessment of the number of orthologous, non-orthologous and unique-paralogous gene copies predicted to encode secreted proteins across the phylum Microsporidia	126
Figure 2.20: Comparing the enzyme repertoire of the pentose phosphate and deoxyribonucleotide synthesis pathway between members and non-members of the Enterocytozoonidae family	144
Figure 3.1: Schematic representation of the glycolytic pathway and alternative routes	151
Figure 3.2: Summary of metabolic pathways retained in the Enterocytozoonidae as revealed by comparative genomic survey	168
Figure 3.3: Loss of glycolytic genes within the Enterocytozoonidae	173
Figure 3.4: Explaining loss of glycolysis in the Microsporidia.....	173
Figure 3.5: Linking glycolysis to the mevalonate pathway	174
Figure 3.6: Phylogeny of hexokinase	176
Figure 3.7: Maximum likelihood analysis performed on a masked alignment of microsporidian hexokinases (HKs).....	177

Figure 3.8: Maximum likelihood analysis performed on a masked alignment of microsporidian PTPA proteins	178
Figure 3.9: Comparing gene order conservation between hexokinase-containing contig in <i>Enterospora canceri</i> and corresponding contigs within the <i>Enterocytozoon hepatopenaei</i> , <i>Hepatospora eriocheir</i> and <i>Vittaforma corneae</i>	179
Figure 3.10: Comparing gene order conservation between hexokinase-containing contig in <i>Vittaforma corneae</i> and corresponding contigs within the Enterocytozoonidae.....	180
Figure 3.11: Understanding the evolution of Chimeric hexokinase in <i>Enterospora canceri</i>	181
Figure 3.12: Putative chimeric proteins in the genome of extant microsporidians	181
Figure 3.13: Maximum likelihood analysis performed on a masked alignment of microsporidian ATP/ADP translocases	183
Figure 3.14: Western blot analysis of microsporidian hexokinases expressed in <i>Escherichia coli</i>	185
Figure 3.15: NADH absorbance readings for hexokinase/glucose-6-phosphate coupled reaction	187
Figure 3.16: BIOLOG Phenotypic microarray analysis of Rosetta(DE3)pLysS <i>E. coli</i> cells transformed with hexokinases from three different microsporidian species	188
Figure 4.1: Separate maximum likelihood trees of 20 microsporidians.....	212
Figure 4.2: Phylogenetic trees based on A. Maximum likelihood B. Bayesian inference of 20 microsporidians for six concatenated genes rooted with <i>Saccharomyces cerevisiae</i>	213

List of Tables

Table 1.1: List of species within the Enterocytozoonidae family, site of intracellular infection, hosts and host's habitat and host's commercial value	49
Table 2.1 Strain names of <i>Escherichia coli</i> used, molecular features, sources and purpose of use	158
Table 2.2 Strain names of <i>Saccharomyces cerevisiae</i> used, molecular features, sources and purpose of use	158
Table 2.1: Comparing the performance of four eukaryotic genome assembler on the genome of <i>Hepatospora eriocheir canceri</i>	94
Table 2.2: Comparing SPADES assemblies at different Illumina read GC content cut-offs for the genome of <i>Hepatospra eriocheir canceri</i>	95
Table 2.3: Comparing A5-MISEQ assemblies at different Illumina read GC content cut-offs for the genome of <i>Hepatospora eriocheir canceri</i>	95
Table 2.4: Comparing VELVET assemblies at different Illumina read GC content cut-offs for the genome of <i>Hepatospora eriocheir canceri</i>	96
Table 2.5: Comparing RAY assemblies at different Illumina read GC content cut-offs for the genome of <i>Hepatospora eriocheir canceri</i>	96
Table 2.6: Comparing SPADES and A5-MISEQ assemblies performed on the raw Illumina reads of <i>Enterocytozoon hepatopenaei</i>	101
Table 2.7: Comparing SPADES assemblies at different Illumina read GC content cut-offs for the genome of <i>Enterocytozoon hepatopenaei</i>	102
Table 2.8: Comparing A5-MISEQ assemblies performed with Illumina paired-end reads filtered at different GC percentage content of <i>Enterocytozoon hepatopenaei</i>	103
Table 2.9: Statistics for the final assembly of <i>Enterocytozoon hepatopenaei</i>	105
Table 2.10: Comparing SPADES and A5-MISEQ assemblies performed on the raw Illumina reads of <i>Enterospora canceri</i> obtained from the first sequencing run.....	107
Table 2.11: Comparing SPADES and A5-MISEQ assemblies performed on the raw Illumina reads of <i>Enterospora canceri</i> obtained from the second sequencing run.....	107

Table 2.12: Comparing SPADES and A5-MISEQ assemblies performed on the raw Illumina reads of <i>Enterospora canceri</i> obtained from the third sequencing run.....	108
Table 2.13: Statistics for the genome of <i>Enterospora canceri</i> assembled with reads that remapped onto filtered contigs.	111
Table 2.14: Remapping statistics RUN1 and 2	112
Table 2.15: <i>De novo</i> assembly statistics of <i>Enterospora canceri</i> genome assembled with reads from RUN1 and RUN2.....	112
Table 2.16: <i>De novo</i> assembly statistics of <i>Enterospora canceri</i> genome assembled with Illumina reads that mapped onto the filtered RUN3 contigs	114
Table 2.17: <i>De novo</i> assembly statistics of <i>Enterospora canceri</i> genome assembled with reads from RUN1, 2 and 3	114
Table 2.18: Statistics for assembled genomes submitted to GENBANK .	115
Table 2.19: Distribution of predicted plasma membrane transporters not assigned to a protein family	125
Table 2.20: Predicted secreted proteins annotated by MAKER and/or BLAST2GO	127
Table 2.21: Functional annotation of orthoclusters of secreted proteins unique to 2 or more members of the Enterocytozoonidae	129
Table 4.1: Morphological differences between <i>Hepatospora</i> and <i>Hepatospora</i> -like microsporidia.....	206
Table 4.2: Higher sequence similarity between the three <i>Hepatospora</i> / <i>Hepatospora</i> -like species than between strains/subspecies of other microsporidia.....	214

Abbreviations

°C	Degrees Celsius
3'	Three Prime
5'	Five Prime
6PGIte	6-phosphoglucono- δ -lactonate
ADP	Adenosine diphosphate
AMP	Adenosine monophosphate
AOX	Alternative oxidase
ATP	Adenosine triphosphate
BI	Bayesian Inference
BLAST	Basic Local Alignment Search Tool
bp	Base Pair
BWA	Burrows-Wheeler
cAMP	Cyclic adenosine monophosphate
CEFAS	Centre for Environment, Fisheries and Aquaculture Science
CTP	Cytidine triphosphate
DAP	Dihydroxyacetone phosphate
DNA	Deoxyribonucleic acid
dNDP	Deoxyribonucleotides diphosphates
dNMP	Deoxyribonucleotides monophosphates
dNTP	Deoxyribonucleotide
DTT	Dithiothreitol
EC	Enzyme Commission
ED	Entner-Doudoroff
EDTA	Ethylenediaminetetraacetic acid
EEC	European Economic Community
EM	Electron Microscopy
<i>et al.</i>	and others
Fru1,6BP	Fructose 1,6-bisphosphate
Fru6P	Fructose 6-phosphate
g	Grams
g/L	Grams per litre
G13BP	1,3-bisphosphoglycerate
G23BP	2,3-bisphosphoglycerate
G2P	2-phosphoglycerate
G3P	3-phosphoglycerate
GDP	Guanosine diphosphate
GFP	Green fluorescent protein
Glc6P	Glucose 6-phosphate
GLK	Glucokinase
GNAT	Gcn5-related N-acetyltransferases
GO	Gene ontology
GPD	Glyceraldehyde phosphate dehydrogenase
GPI	Glycosylphosphatidylinositol
GTP	Guanosine triphosphate
H+	Hydrogen ion

H ₂ O	Water
H ₂ O ₂	Hydrogen peroxide
HCl	hydrochloric acid
HIS-tag	6xpolyhistidine tagged
HIV	Human Immunodeficiency Virus
HK	Hexokinase
HSP70	70 kilodalton heat shock protein
ITS	Internal transcribed spacer
kb	Kilobase
KCl	Potassium Chloride
K _M	Michaelis constant
KOH	Potassium Hydroxide
L	Litre
LB	Luria broth media
LBA	Long branch attraction
LiCl	Lithium chloride
M	Molar
MALDI-MS	Matrix-assisted laser desorption/ionization-mass spectrometry
MAT	Mating type
Mbp	Megabase pair
mg	Milligram
mg/ml	Milligram per millilitre
MgCl ₂	Magnesium Chloride
ml	Millilitre
ML	Maximum likelihood
ml/L	Millilitre per litre
mM	Millimolar
Na	Sodium
NaCl	Sodium Chloride
NAD ⁺	Nicotinamide adenine dinucleotide
NADH	Reduced nicotinamide adenine dinucleotide
NCBI	National center for biotechnology information
NDP	Nucleoside diphosphate
NGS	Next Generation Sequencing
Ni ²⁺	Nickel ion
nm	Nanometres
NMP	Nucleoside monophosphate
NTA	Nitrilotriacetate
NTP	Nucleoside triphosphate
NUDIX	Nucleoside diphosphate linked to some moiety, X
OD ₆₀₀	Optical density at 600 nm
OLC	Overlap-Layout-Consensus
ORF	Open reading frame
PACBIO	Pacific Biosciences
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PEG	Polyethylene glycol

PEP	Phosphoenolpyruvate
pH	Measure of acidity or alkalinity of a solution (concentration of protons in a solution)
PTPA	Protein-tyrosine-phosphatase
R5P	Ribose 5-phosphate
rDNA	Ribosomal deoxyribonucleic acid
Rib5P	Ribulose 5-phosphate
RNA	Ribonucleic acid
ROS	Reactive oxygen species
SD	Synthetic Defined
SDS	Sodium dodecyl sulphate
SGA	Synthetic genetic array
SGD	<i>Saccharomyces</i> genome database
SGS	Slow Growth Syndrome
TAE	Tris-acetate-EDTA
TE	Transposable Element
TEM	Transmission electron microscopy
tRNA	Transfer RNA
UDP	Uridine diphosphate
UK	United Kingdom
UTP	Uridine triphosphate
V	Volts
VDAC	Voltage-dependent anion channels
V_{max}	Maximal velocity
WDS	Winter Disease Syndrome
WGS	White Faeces Syndrome
X5P	Xylulose 5-phosphate
YPD	Yeast extract peptone media
μg	Micrograms
$\mu\text{g/ml}$	Micrograms per millilitre
μl	Microlitre
μM	Micromolar

Chapter 1 General Introduction

1.1 The Microsporidia: structure and development

Microsporidians are unicellular eukaryotic organisms but lack canonical mitochondria and centrioles (Vávra and Larsson 2014). These opportunistic obligate intracellular parasites belong to the phylum Microsporidia and infect most animal lineages (Sprague 1977; Sprague & Becnel 1998; Vávra & Larson 2014). They have a complex life cycle and the exact cycle stages varies between species (Vávra and Larsson 2014). However, the microsporidian life cycle can be summarised into three main stages: A proliferative phase that only develops within the host cell, a sporogonial phase that produces infective entities and the infective spore phase (Wittner and Weiss 1999).

1.1.1 The spore

The spore is a quiescent life stage which is encased in a double layered chitinaceous cell wall approximately 0.1 μm in thickness (Kudo 1921; Vávra 1977; Vávra & Larson 1999) (Figure. 1.1). On TEM, the outer cell wall of the genus *Encephalitozoon* has been shown to consist of an electro-lucent middle layer sandwiched between a spiny outer and a fibrous inner chitinaceous layer (Bigliardi et al. 1996; Bigliardi et al. 1997). Thus the plasma membrane is enveloped in at least three chitinaceous fibrous reinforcements, which protects the cell from external factors such as desiccation and toxins (Vávra 1977; Bigliardi & Sacchi 2001). Also, the outer layer or exospore of some species is ornamented with villous protrusions, fibrous filaments, tubular mesh, tail like projections or a mucus layer (Larsson 1989; Overstreet & Weidner 1974; Vávra 1977; Rausch & Grunewald 1980) (Figure 1.1). These spore appendages are presumed to aid in host transmission and anchoring (Overstreet & Weidner 1974; Stentiford et al. 2013). The mucus layer, which is often observed in aquatic species, has also been speculated to aid in buoyancy (Vávra 1977; Vávra & Larsson 2014).

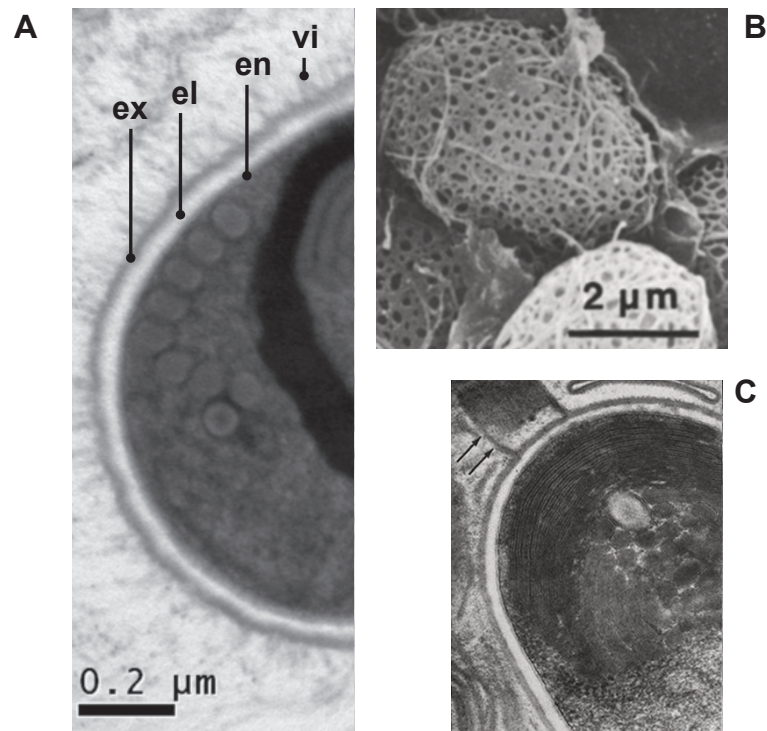


Figure 1.1: Polymorphic nature of the microsporidian spore wall. A. Spore ultrastructure of *Ameson pulvis* displaying villous protrusions on spore wall; vi, endospore; en, electrolucent middle layer of spore wall; el, and exospore; ex. Modified from (Stentiford et al. 2013). B. Scanning electron micrograph of *Pleistophora simulii* spore. Exospore characterized by a network of fine thread-like structures. From (Rausch & Grunewald 1980). C. Ultrastructure of *Inodosporus spraguei* (*Thelohania spraguei*) displaying tail-like structure attached to the apical end of its exospore. (Magnification=X61600) (From Overstreet & Weidner 1974)

The size of microsporidian spores varies between species with shapes including spherical, pyriform and bacillus. However, different spore shapes and sizes have also been recorded for the same species (Stentiford et al. 2013; Vávra & Luke 2013). Under the microscope, the microsporidian spore is refractory and ranges between 1-12 μm in size (Kudo 1930; Freeman et al. 2004; Cali et al. 1993). Ultrastructural features of the spore have not only been pivotal in making the Microsporidia a distinct *bona fide* phylum but it has also been utilized in the systematics within the phylum (Vávra & Larsson 2014). One of these features is the extrusion apparatus, which is arguably the definitive feature of the Microsporidia. It is composed of an electron dense polar tube (polar filament) that forms 4-30 coils within the sporoplasm and anchors itself at the anterior end of the plasma membrane with a polar sac-anchoring disc complex, a structure rich in glycoproteins (Kudo 1920; Vávra 1977; Taupin et al. 2007) (Figure 1.2). The number of coils formed by the polar tube varies between species although it is estimated that about 20 % of known Microsporidian species possess an uncoiled version of this apparatus (Canning & Vávra 2000). The polar sac-anchoring disc

complex covers the proximal part of the polaroplast, a single-membrane tubular mesh. Together with the posterior vacuole, the polaroplast plays an active role during spore germination.

The trigger for spore germination varies between microsporidian species (Undeen & Epsky 1990). For example, *Vavraia culicis* spores have been reported to germinate *in vitro* in alkaline environments of pH 7.0-9.0 (Undeen 1983) whereas germination in of *Nosema locustae* spores occurs by an initial dehydration followed by a rehydration step (Whitlock & Johnson 1990). These germination triggers initiate a signalling cascade within spores that culminate in the extrusion of the polar filament (Figure 1.2). However, the exact steps in this cascade are still unknown and it is thought that it may be dependent on habitat. For example, enzymes such as trehalase, that have been shown to play a pivotal role in polar filament extrusion, have been detected mostly in aquatic and not in terrestrial microsporidians (Undeen & Vander Meer 1999). Catalase activity within the posterior vacuole has also been linked to polar filament germination (Findley et al. 2005) although there is no evidence for catalase in microsporidian genomes that could perform this role (Fast et al. 2003; Williams et al. 2014).

Even though the germination triggers and signalling cascades may vary between microsporidian lineages, they all result in the establishment of an intrasporal osmotic gradient that in turn causes an influx of water from the environment into the sporoplasm, possibly by the aid of dedicated water channels called aquaporins (Frixione et al. 1992; Frixione et al. 1997; Ghosh et al. 2006). The surge in intrasporal pressure, which is postulated to be around 79 atmospheres, resulting from the rapid water influx culminates in the extrusion of the polar filament and the swelling of the posterior vacuole (Undeen & Vander Meer 1994) (Figure 1.2). The polar filament breaks through the spore wall and plunges into a surrounding host cell, piercing the host cell membrane (Cali et al. 2002; Takvorian et al. 2005) (Figure 1.2). The swollen posterior vacuole pushes the spore contents (nucleus and ribosome dense sporoplasm) through the everted polar filament into the host cell leaving behind an empty spore. The spore contents enter the host as a single-membrane cell which initiates the next life cycle (Vávra & Larson 2014). There is evidence suggesting that during spore germination, the polaroplast contributes to building spore turgor needed to force the polar filament out of the spore (Weidner et al. 1984). Once in the host cell, the polaroplast

membrane is thought to form the initial plasma membrane of the first meront (Weidner et al. 1984).

1.1.2 The meront

The expulsion of the sporoplasm into the host cell marks the beginning of a new life stage, the meront. This is a single membrane bound cell usually found in direct contact with the host cytoplasm that utilises host resources to replicate into a cluster of meronts, often leading to host cell and nuclear enlargement (Liu, 1972). Merozoites are rich in ribosomes and have enlarged nuclei as compared to other developmental stages, perhaps as a result of increased protein demand for replication (Vávra & Larson 1999).

1.1.3 The sporont

Towards the end of merogony, each daughter meront develops into a sporont. Sporonts have smaller nuclei, fewer cytoplasmic ribosomes and more endoplasmic reticulum as compared to meronts and also have a layer of electron dense material on their plasma membrane. The appearance of this electron dense material marks the end of merogony and the commencement of the next developmental stage, sporogony. This thick electron dense material will eventually mature into the thick outer layer of the exospore (Vávra & Larson 1999). In some microsporidian species, sporonts also undergo a number of replications before forming sporoblasts, which later develop into mature spores. Spore-specific structures such as polaroplasts (Takvorian & Cali 1994) and polar filaments become visible under the microscope in the sporoblast (Larsson 1994) although the posterior vacuole develops towards the end of sporogony (Sprague & Vernick 1969). The mature sporont begins to deposit about 100 nm of electron-lucent material which would later develop into the chitinaceous endospore (Vávra et al. 1986).

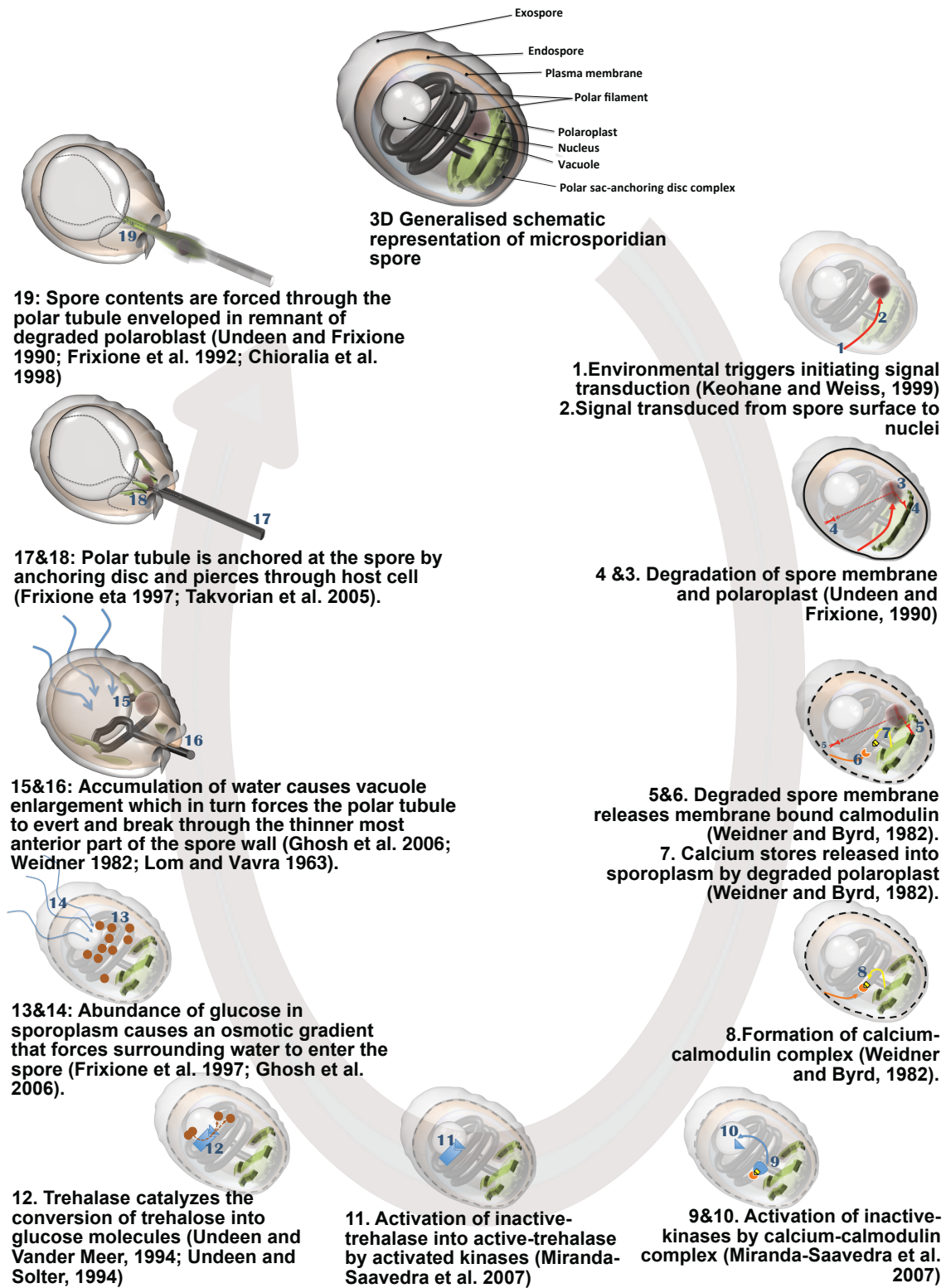


Figure 1.2: 3D cartoon of a microsporidian spore: Summarizing the different stages (1-19) involved in polar filament extrusion.

1.2 History of the phylogenetic placement of the Microsporidia

1.2.1 Coining of the name Microsporidia and addition to the Schizomycetes

In the middle of the nineteenth century the European silk industry was at the brink of complete collapse as a result of a then unknown disease which caused mortality in silk worm larvae. The pébrine disease, as it was called, was eventually linked to 'Yeast-like' fungal agents by Swiss microbiologist Karl Wilhelm von Nägeli in 1857. He named them *Nosema bombycis* and considered them to be members of the schizomycete fungi. Members of this currently disbanded group included both yeast and bacteria (reviewed in Corradi & Keeling 2009). Following this, French embryologist Édouard-Gérard Balbiani added *Nosema* to the Sporozoa group in 1882. At the time, the Sporozoa consisted of a consortium of distantly related spore forming alveolates, rhizarians, animals, green algae and protists of unknown origins. Although Balbiani's classification did not reflect phylogenetic history, he created a new group for *Nosema* and named it Microsporidia - a name used for this phylum to date. Balbiani's work was originally published in French in (Balbiani 1882) but recently reviewed in (Corradi & Keeling 2009).

1.2.2 The Archezoa hypothesis

From the late 1950s, electron microscopy studies begun to suggest that microsporidians were amitochondriate, and lacked peroxisomes and Golgi bodies. Albeit fascinating at the time, these findings led researchers to erroneously classify the Microsporidia as Archezoa, primitive eukaryotes that diverged before the symbiotic event that led to mitochondria in extant eukaryotes (Kudo & Daniels 1963; Vávra 1965; Cavalier-Smith 1983). Indeed, the Kingdom Archezoa is also now disbanded and its members, Archamoebae, Parabasalia, Metamonada and Microsporidia are also now considered to be eukaryotic lineages that have undergone secondary mitochondrial reduction (Clark & Roger 1995; Bui et al. 1996; Roger et al. 1996; Hinkle et al. 1994; Morin & Mignot 1995). Nevertheless, there were some compelling findings at the time that advocated the creation of the Archezoa Kingdom. For instance, biochemical studies showed ribosomes of the Microsporidia and other members of the Archezoa to possess a sedimentation coefficient of 70S, a characteristic feature of prokaryotes (Ishihara & Hayashi 1968; Cavalier-Smith 1987). Microsporidian 5.8S and 23S

rDNAs also seemed to have a fused conformation, another characteristic feature of prokaryotes (Vossbrinck & Woese 1986; Vossbrinck et al. 1987). Moreover, the renaissance of phylogenetic studies based on molecular data in the 1980s provided even more evidence to support the Archezoa hypothesis. Both studies based on rDNA and protein coding genes such as elongation factor 1 alpha and elongation factor 2 seemed to place members of the Archezoa at the base of the eukaryotic tree of life (Cavalier-Smith 1983; Vossbrinck et al. 1987; Kamaishi, Hashimoto, Nakamura, Nakamura, et al. 1996; Kamaishi, Hashimoto, Nakamura, Masuda, et al. 1996).

1.2.3 The mitosome and the end of the Archezoa hypothesis

In the subsequent decades, evidence against the Archezoa hypothesis began to gradually emerge. In summary, these studies showed that aggregation of the so-called Archezoa at the base of the eukaryotic tree was as a result of the accelerated evolution characteristic of the genomes of members of this group and the inability of both parsimony and maximum likelihood-based phylogenetic algorithms at the time to take discrete changes in site rate distribution into account. This consequently culminated in the artefactual clustering of fast evolving species at the base of the eukaryotic tree, a phenomenon known as Long Branch Attraction (LBA) (Felsenstein 1978; Stiller & Hall 1999; Inagaki et al. 2004). The mid-1990's saw the development of more powerful phylogenetic tools that were able to take discrete changes of site rate variation into account and the use of increased numbers of protein coding genes for microsporidian phylogenetic analysis. One of the first phylogenetic studies to provide evidence for the fungal relationship of the Microsporidia was a study by Keeling and Doolittle, 1996. Their phylogeny, based on alpha and beta-tubulin genes, suggested the placement of the Microsporidia at the base of the fungal tree (Keeling & Doolittle 1996). A flurry of studies based on single proteins, all with the aim of understanding the phylogenetic placement of the Microsporidia have been produced since then. Some of these proteins include the mitochondrial HSP70 (Germot et al. 1997; Hirt et al. 1997), RNA polymerase II (Hirt et al. 1999) TATA-box binding protein (Fast et al. 1999) and glutamyl synthase (Brown & Doolittle 1999). Similar to Keeling and Doolittle's results, these studies suggested a fungal relationship of the Microsporidia making the Archezoa hypothesis increasingly unpopular. Since the premise that Microsporidia were Archezoa was

originally based on the absence of mitochondria, the discovery of mitochondrial derived genes in the nuclear genomes of several microsporidian species [*HSP70* (Germot et al. 1997; Hirt et al. 1997; Peyretailade et al. 1998), pyruvate dehydrogenase alpha and beta E1 (Fast & Keeling 2001), *ATM1*, *ISU1/ISU2*, *NFS1*, *SSQ1*, *YAH1* and *PDB1* (Katinka et al. 2001)] further decreased credibility in the Archezoa hypothesis. Perhaps what ultimately put to rest any notion that Microsporidia were Archezoa was the immunolocalization of mitochondrial HSP70 to what is now considered as a relict microsporidian mitochondrion, the mitosome (Williams et al. 2002) and also the discovery of fungi-like features in the Microsporidia. Some of these features are the presence of meiotic and closed mitotic cycles, chitin and trehalose metabolic genes (Keeling & Doolittle 1996; Germot et al. 1997) and the presence of an 11-12 amino acid insertion within the elongation expression factor I gene of *Glugea plecoglossi* which was thought to be exclusive to animals and fungi (Kamaishi et al. 1996).

1.2.4 The phylogenetic link between the Microsporidia and Fungi

As such, whether the Microsporidia diverged outside or within the fungal tree was still a contentious topic by the late 1990s. In 2000, Keeling *et al.* (2000) performed a phylogenetic study with alpha and beta-tubulin genes from representatives of four fungal phyla and five microsporidian species. Even though this study was unable to conclusively show which fungal lineage the Microsporidia diverged from, it suggested that the diversification of the microsporidian lineage may have happened after the divergence of the Chytridiomycota and that they may possibly be a sister group to the Dikarya clade (Keeling et al. 2000) (Figure 1.3). Results from follow up studies went on to suggest that Microsporidia may have shared an ancestor with extant zygomycetes (Keeling 2003) (Figure 1.3). Again, these results were far from conclusive as the author admittedly wrote that alternative tree topologies where the Microsporidia branched after Chytrids but before all other fungi could not be statistically rejected (Keeling 2003). It was however becoming increasingly evident at this stage that the tubulin genes were prone to an accelerated rate of nucleotide substitution in the Dikarya and Microsporidia and hence were unsuitable as a phylogenetic marker as they led to long branch attraction and artefacts in trees (Keeling 2003; Gill & Fast 2006). Authors at this point had also begun to look at the use of multiple loci in assessing the fungi-Microsporidia phylogenetic relationship (Keeling 2003). With the increased

availability of draft genomes, Gill and Fast (2006) approached this problem by taking a multi-protein [alpha-tubulin, beta-tubulin, the largest subunit of RNA polymerase II (RPB1), DNA repair helicase RAD25, TATA-box binding protein (TBP), subunit of the E2 ubiquitin conjugating enzyme (UBC2), and alpha and beta subunits of pyruvate dehydrogenase E1] concatenated approach to assess the position of Microsporidia in the fungal tree. In their results, the Microsporidia branched as a sister group to the Ascomycetes and Basidiomycetes (Figure 1.3). In the same year, a similar study based on a different set of both protein coding and non-coding genes (28S rDNA, 5.8S rDNA, elongation factor-1 (*EF-1*), *RPB1* and *RPB2*) suggested microsporidians to be closely related to fungal parasites, *Rozella*, and branch at the base of the fungal tree (James et al. 2006) (Figure 1.3). Contrary to expectations, this concatenated protein approach also failed to conclusively resolve the microsporidian-fungi relationship. Even though Gill and Fast (2006) and James *et al.* (2006) placed the Microsporidia in contrasting positions with relation to the fungal tree, both authors could not statistically reject the alternative phylogenetic placements. Some authors at this stage had begun exploiting different genomic information such as synteny to address this problem. The observation that Microsporidia and zygomycetes share gene order conservation at their MAT sex locus, led to the suggestion that the microsporidia were related to the zygomycetes (Lee et al. 2008). However, Koestler and Ebersberger (2011) and Capella-Gutierrez *et al.* (2012) challenged Lee's findings by demonstrating that synteny across whole genomes is not significantly higher between microsporidians and Zygomycota than between microsporidians and other fungal groups. Just as the study performed by James *et al.* (2006), Capella-Gutierrez *et al.*'s work which benefitted from an expansive set of sampled microsporidian species and a total of 53 concatenated orthologous proteins supported the branching of the Microsporidia at the base of the fungal tree with high statistical support. Capella-Gutierrez *et al.* demonstrated that statistical support for alternate hypothesis decreases with the removal of highly variable sites from the alignment (Capella-gutiérrez et al. 2012). This was also observed by James *et al.* (2013) in a phylogenomic study involving 200 concatenated orthologous protein sets from microsporidian and fungal species including *Rozella allomycis*, a parasitic fungi previously suggested to be the closest relative of Microsporidia albeit with poor statistical support (James et al. 2006; Karpov et al. 2013). In this study the relationship between the Microsporidia and *Rozella*

allomycis was strongly supported by statistics and these groups together formed a sister group to the rest of the fungal lineage at the base of the fungal tree of life (James et al. 2013) (Figure 1.3). As it stands, our current knowledge about diversity at the base of the fungal tree of life is very poor. This was evident when Jones *et al.* (2011) used deep sequencing to identify a plethora of previously unknown *Rozella*-related organisms from environmental samples (Jones et al. 2011). The authors later went on to call this group the Cryptomycota (syn. Rozellomycota) (Richards et al. 2011) (Figure 1.3). Unknown to the authors at the time, a subgroup of the environmental sample sequences from Jones *et al.* (2011) belonged to the Aphelida clade, a group of parasites of algae (Letcher et al. 2013) and current data points towards a stronger relationship between *Aphelida* and Microsporidia than between *Rozella* and Microsporidia (Keeling 2014) (Figure 1.3). Some current data suggests Microsporidia belongs to the newly created Cryptomycota group situated at the base of the fungal tree of life, and together these form a sister group to the rest of the Fungi (Figure 1.3). However, more full genomes from the Cryptomycota are needed to better establish the relationships between these early branching fungal groups.

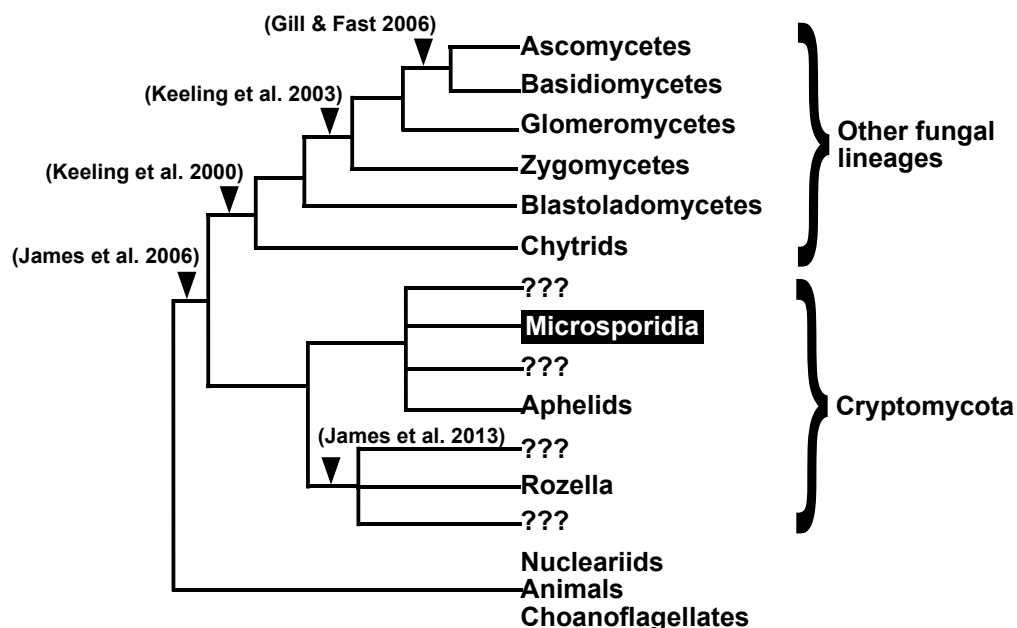


Figure 1.3: Schematic representation of the current view of the phylogenetic position of the Microsporidia. Tree shows the microsporidian phylum to be nested within the Cryptomycota, which has several unknown lineages (???). Triangles represent where past authors have placed the microsporidian lineage. Image adapted from Keeling 2014.

1.3 Systematics within the Microsporidia

1.3.1 Creation of the order Microsporides

Until now, placement of the Microsporidia in the tree of life has certainly dominated research focus as compared to investigations concerning systematics within the phylum itself. Regardless, systematics within the Microsporidia has faced a similar fate of controversy. Arguably it could be said that Balbiani (1882) initiated debate for systematics within the Microsporidia when he assigned the group a taxonomic ranking: order Microsporides. By this time, *Nosema bombycis* was the only member of this group. Thelohan (1892) attempted classifying the known microsporidians at the time and created three genera based on two simple characters: number of spores within sporoblasts (*Thelohania*) and presence of a pansporoblastic membrane (*Glugea* and *Pleistophora*). The use of the pansporoblastic membrane as a taxonomic character would continue to be a delimitating factor in future classification attempts. Interestingly, despite his simplistic classification system, Thelohan's efforts caused confusion for future authors (Gurley 1893; Labbé 1899; Kudo 1924; Sprague 1977) as his newly assigned genera were placed under the order Myxosporida instead of Microsporidia. Myxosporidians are obligate endoparasitic cnidarians whose taxonomic placement at the time was also being debated (Fook & Siddall 2015).

1.3.2 The use of phenotypic traits for taxonomic ranking in the Microsporidia

According to Sprague's assessment of systematics within the Microsporidia in 1977, Thelohan violated "The Law of Priority" by treating *Nosema* (Naegeli, 1857) and *Glugea* (Thelohan, 1891) as synonymous (Sprague 1977).

In his Dutch publication, Labbé (1899) addressed the latter of Thelohan's errors by changing the Glugeidae to Nosematidae family (Labbé 1899). Drawing from this work and that of Pérez (1905) which distinguished between *Glugea* and *Nosema* (Pérez 1905), Stempbell (1909) suggested the creation of the family Glugeidae (Thélohan, 1982) and a completely new family, Pleistophoridae, in addition to the Nosematidae suggested by Labbé (1899) [reviewed in (Sprague 1977)]. In his work, Stempell (1909) put forward a system of classification that used difference in merogonial features as a character for ranking at the family level, sporogonial features as generic ranking and spore features as species ranking (Stempell 1909). Léger and Hesse (1922) also proposed a similar

classification system but relied solely on spore phenotype as a character for ranking. The error of Stempell, Léger and Hesse's classification systems lay in the implied assumption that the above mentioned characters evolved at the same speed across all microsporidian lineages (Simpson 1961).

1.3.3 The use of ultrastructural features for taxonomic ranking in the Microsporidia

With the invention of the electron microscope in the 1930s it was not long until microsporidiologists began to appreciate the depth of diversity they faced and hence commenced using more detailed ultrastructural characters in microsporidian systematics. Thus electron microscopy brought with it a whole range of ultrastructural information that was inaccessible to previous authors and efforts at this time were focused at integrating this new information into already established classification systems. By the mid 1970s there appeared to be two distinct microsporidian groups in the 700 known species (Vávra 1976a): one group with developed (coiled) polar filaments and another with primitive (uncoiled) polar filament and sometimes absent organelles (Sprague 1977). Sprague (1977) used presence or absence of a polaroplast and posterior vacuole to group the Microsporidia into class Microsporea (higher microsporidians) and Rudimicrosporea (primitive microsporidians) respectively. Whereas Rudimicrosporea contained the sole order of Metchnikovellida, the Microsporea were subdivided into two orders, Microsporida and Chytridopsida. The delimitating characters for this taxonomic ranking were polar filament arrangement and spore organelle development. Thus, species that had primitive-looking polar filaments and organelles were grouped in the Chytridopsida order whereas species with coiled polar filaments and well developed organelles fell into the Microsporida (Sprague 1977). Interestingly, even in this new era of electron microscopy, Sprague used presence and absence of pansporoblastic membrane as a delimitation character for suborder taxonomic ranking and characters such as karyotype and sporogony for family taxonomic ranking.

1.3.4 The phylum Microsporidia

Perhaps the most important contribution of Sprague's work to date was to lift the Microsporida from the lower rank of class suggested by previous authors (Tuzet et al. 1971) to the taxonomic ranking of phylum thereby creating more room for the addition of new species. In his work he calls previous classification attempts

“rigid” systems and refers to his work as “open and flexible, thus capable of being infinitely expanded as new taxa become delimited” (Sprague 1977). Indeed, Sprague’s work became the foundation for subsequent studies in microsporidian systematics (Vivier 1979; Corliss 1994; Voronin 2001) but did not fail to attract controversy. In Weiser’s work released only a few months later, the order Chytridopsida was placed under primitive microsporidians (Class: Metchnikovellidea) as opposed to higher microsporidians suggested by Sprague (1977) (Weiser 1977). Surprisingly, there are still no rDNA sequences publicly available for members of the Metchnikovellidea (Weiser, 1977) and the position of this seemingly primitive lineage on protein coding DNA-based phylogenetic trees is yet to be resolved. Based on Sprague’s work and EM-derived ultrastructural information Larson published the first classification system that used number of polar filament coils as a delimitation factor with the aim of resolving lower taxonomic ranks in the Microsporidia (Larsson 1986). By now, there was growing debate of the significance of karyotype as a classification character. Whereas Larsson (1986) had commented about the insignificance of karyotype in microsporidian systematics, Sprague *et al.* (1992) published a new taxonomy guideline heavily based on karyotype.

1.3.5 Ribosomal DNA and systematics of the Microsporidia

The next decade saw the reassignment of many species that were originally ranked by traditional methods to different higher and lower taxonomic ranks due to the emergence of the molecular era. The molecular era occurred as a cumulative consequence of increased computational power and availability of rDNA sequences. rDNA-based phylogenetic trees began to reveal that species of *Varimorpha* and *Nosema*, clumped in the same family of Nosematidae, were not as closely related as suggested by their diplokaryotic spores and their pansporoblastic membrane-covered sporoblasts as hypothesised by Weiser (1977) (Baker *et al.* 1994; Müller *et al.* 2000). Following this, phylogenetic evidence based on rDNA was used to reassign *Nosema locustae* to the genus *Paranosema* (Sokolova *et al.* 2003) and to *Antonospora* the following year (Slamovits, *et al.* 2004). *Endoreticulatus eriocheir* (Wang & Chen 2007) was reassigned as *Hepatospora eriocheir* (Stentiford *et al.* 2011). Stentiford *et al.* (2010) proposed the erection of higher taxonomic ranking (Family: Myosporidae and Order: Crustaceacida) to accommodate a new species that would have been

assigned to the Thelohanidae family if traditional morphological criteria were employed. In a similar rDNA-based study *Thelohania butleri*, a freshwater parasite which would have been added to the Thelohanidae family based on morphological criteria, showed close phylogenetic affinity to an unknown microsporidian which belonged to the marine clade (Brown & Adamson 2006). Thus *Thelohania butleri* would have been misplaced at both high and low taxonomic levels if traditional criteria were employed.

It was evident at this point that the microsporidian phylum needed to be revised with the guidance of the increasingly available molecular evidence. To this end, Vossbrinck and Vossbrinck-Debrunner (2005) provided the first expansive phylogenetic analysis based on molecular characters (rDNA) for 125 species.

In their findings, it became evident that habitat rather than ultrastructural features better united microsporidian lineages at higher taxonomic levels. However, Vossbrinck and Debrunner-Vossbrinck's work lacked sample taxa from the Metchnikovellidae (Weiser, 1977). Consequently, the question of whether the short and under developed polar filament of members of this group was as a result of their primitive nature or a product of host-specialization still remained unanswered. To summarise, systematics within the Microsporidia based on traditional methods have been shown to be unreliable for assignment of taxonomic ranks in the Microsporidia (Stentiford et al. 2013). This is because physical characters used evolve too rapidly making them unsuitable for delimitation especially for higher taxonomic ranks such as subclasses and orders (Vossbrinck & Debrunner-Vossbrinck 2005). In recent years, independent research groups have reassessed individual lineages including the Nosematidae, Thelohanidae and Enterocytozoonidae families with member species reassigned or added. The only higher taxonomic assignments that seem to have stood the test of time are the Phylum: Microsporidia (Sprague 1977) and the subclasses Aquasporidia, Marinosporidia, Terresporidia suggested by Vossbrinck and Debrunner-Vossbrinck (2005). With DNA sequencing becoming increasingly cheaper, a taxonomic reassessment based on rDNA or other markers of the 1500 species currently known is feasible but such a venture will require a tremendous international collaborative effort.

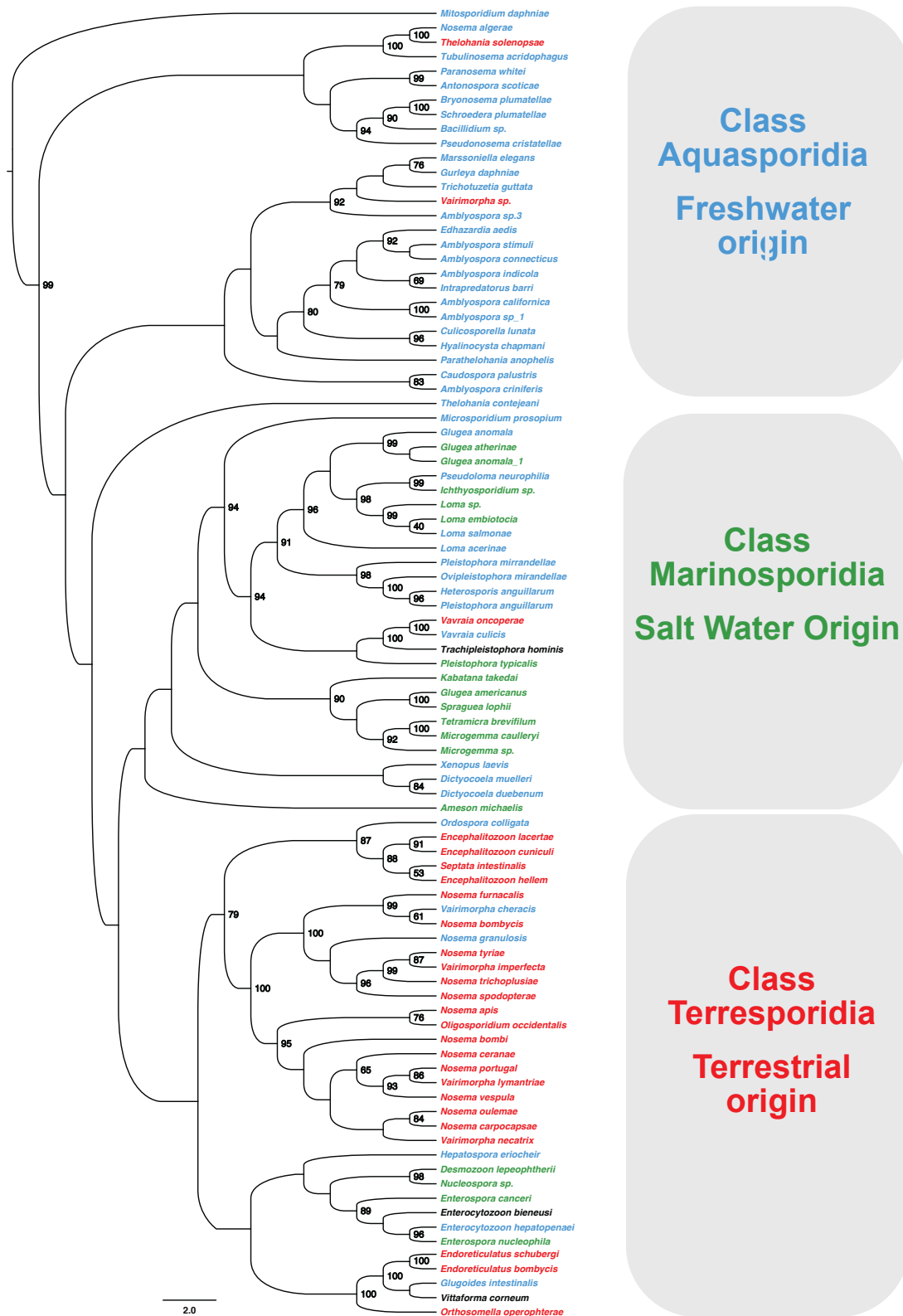


Figure 1.4: rDNA-based phylogeny of 92 sampled microsporidian species demonstrates that microsporidian lineages are better united by habitat-Freshwater (Blue font), Marine (Green font) and Terrestrial (Red font) habitats. The initial alignment and masking of rDNA sequences was performed with command line tool MUSCLE (Edgar 2004) and TRIMAL (Capella-Gutierrez et al. 2009) respectively. Final maximum likelihood analysis was performed with RAXML (Stamatakis 2014) and bootstrap confidence levels >75 % from 100 replicates are displayed on the nodes of the tree. Sequences used were sampled from Vossbrinck and Debrunner (2005) and more recently described lineages: *Mitosporidium daphniae*; XM_013382851.1, *Hepatospora eriocheir*, *Enterocytozoon hepatopenaei*; KF362130.1, *Enterospora nucleophila*; KF135645.1,

1.4 The reduced microsporidian genome and intracellular living

1.4.1 Obligate intracellular parasitism: A consequence of an ancient facultative relationship

It is hypothesised that obligate intracellular organisms may have arisen by initially acquiring a facultative residence within host cells that later progressed to an obligate intracellular lifestyle. This initial facultative relationship then progressed into an endosymbiosis where both the host and the invading organism benefit from each other (Casadevall 2008). A classic example of this is the intimate relationship between the eukaryotic nucleus and mitochondria which began as a symbiotic relationship between a primitive eukaryotic cell and an aerobic bacterium (Margulis 1970; Margulis 1981; Gray et al. 1999). However, not all primitive facultative relationships between a host and an invading organism develop into a symbiosis. Some could also progress into a parasitic relationship (Andersson & Kurland 1998). Unlike symbiosis, parasitism constitutes a continuous tug-of-war between the host's defence mechanisms and the parasite's exploitation system (Andersson & Kurland 1998). Very little is known about the factors that tip an initial facultative relationship towards an endosymbiosis or parasitism but it has been suggested that the source of the initial infection may play a role (Casadevall 2008). Regardless of whether an initial facultative relationship between two organisms takes a trajectory of endosymbiosis or parasitism, intracellular living is hallmarked by reduced genome size in both prokaryotes and eukaryotes and microsporidians are no exception (Katinka et al. 2001; Fraser-Liggett 2005; Corradi et al. 2009; Corradi et al. 2010; Cuomo et al. 2012; Pombert et al. 2012; Heinz et al. 2012; Pan et al. 2013; Campbell et al. 2013; Pombert et al. 2013; Haag et al. 2014; Pombert et al. 2015). In fact, the genome of *Encephalitozoon intestinalis* (Corradi et al. 2010) represents the smallest eukaryotic genome, averaging around 2.3 Mbp, making it even smaller than genomes of some intracellular prokaryotes (Amaro et al. 2012).

1.4.2 The evolutionary trajectory of intracellular lineages is dictated by cell cycle and host environment

Although reduced genomes are a common trait exhibited by intracellular organisms, it is important to highlight that this trait is as a consequence of different evolutionary pressures acting on each intracellular lineage independently. That is, the biochemical rapport between different intracellular lineages and their hosts vary considerably and hence a combination of different evolutionary pressures will act on different lineages. For instance, *Wolbachia* spp. undergo their entire life cycle within their arthropod host cells and are vertically transmitted during host cell replication. As such, they rely entirely on their host for replication (Werren et al. 2008). Since *Wolbachia* spp. only have contact with the host cell cytoplasm, evolutionary pressures here would be different to those in organisms that have an extracellular stage in their lifecycle.

The parasitic intracellular protists, *Leishmania* spp. and *Toxoplasma gondii* both replicate within mammalian macrophages primarily to evade host's immunity and to disseminate across the host's body (Wiser 2011). These organisms rely on the intracellular environment predominantly for host immune system evasion and dissemination. Despite the similarities between these two parasites, a big distinction can be made between the forces driving their evolutionary trajectory: *T. gondii* has an environmental stage that is in direct contact with soil where competition for nutrients and threat of predation and desiccation is high as compared to the comparatively aseptic environment enjoyed by *Leishmania* whose life cycle alternates solely between the blood stream of the mammalian host and the gut of the arthropod host (Wiser 2011). As such, despite the evolutionary pressures for genome reduction acting on these protists due to their intracellular lifestyles, there would also be other evolutionary pressures to maintain and possibly gain genetic/genomic material to survive the extracellular stage of their life cycle.

1.4.3 Factors contributing to genome shrinkage in intracellular parasites

The uniqueness in biochemical interactions between intracellular parasites and their respective hosts make it difficult to find unifying factors responsible for genome reduction in intracellular parasites. However, the recent boom in genomic knowledge as a result of cheap Next Generation Sequencing is making it increasingly evident that unifying evolutionary factors responsible for genome

compaction in intracellular organisms indeed do exist. Some of these factors have been outlined below with a specific focus on the Microsporidia.

1.4.3.1 Cell size

According to Cavalier-Smith (2005), there is a selective pressure for parasites to assume a smaller cell size as this increases overall parasitic efficiency. For instance, small parasites can take refuge in host cells thereby evading the innate immune system. Cell size reduction is also accompanied by nuclear volume reduction—a phenomenon known as karyoplastic ratio maintenance (Cavalier-Smith 2005), where a selective pressure for deletion mutations results in reduced genome size (Figure 1.5) (Andersson & Kurland 1995; Andersson & Kurland 1998).

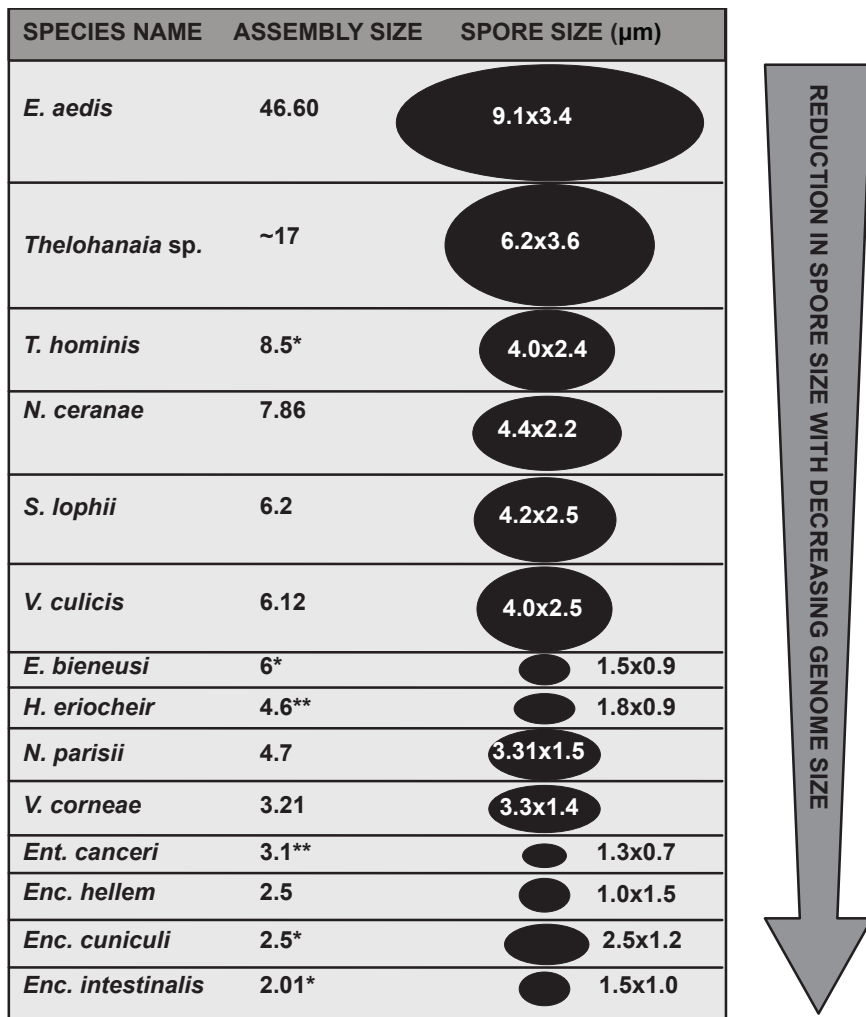


Figure 1.5: Relationship between genome size and relative spore size for selected microsporidian species. *Completely sequenced genome. **Genome sequenced in this study. Black circles represent relative spore size. Pearson correlation statistical analysis gave a value of 0.8992. This suggests a strong positive correlation between spore size and genome/assembly size.

1.4.3.2 Metabolic Economy

The ability to efficiently utilise external resources for personal benefit is required across the spectrum of living organisms. Reliance on external resources, however, is often accompanied by the loss of an organism's endogenous ability to produce such resources. The exogenous availability of a resource lightens the selective pressure to maintain the integrity of the gene or genes responsible for synthesising this resource endogenously. This leads to the accumulation of non-functional homologs due to deleterious mutations and an eventual complete decay of genes, contributing to reduced genome size. Typical examples of this include obligate bacterial endosymbionts of insects such as *Blochmannia*, *Buchnera* and *Wigglesworthia* who have a reduced genome size as compared to their free-living counterparts due to loss of metabolic genes (Wernegreen et al. 2002; Charles & Ishikawa 1999; Akman & Aksoy 2001). A similar case is seen in obligate intracellular parasites, *Rickettsia* spp. who have also lost many biosynthetic metabolic pathways and even show similar functional profiles to the "enslaved" eukaryotic mitochondria (Andersson et al. 1998; Andersson & Andersson 1999; Renesto et al. 2005). This phenomenon is also seen across the entire microsporidian phylum where metabolic capabilities have been lost, reducing their genomes to approximately 2,500 genes (Aurrecochea et al. 2011). This delegation of some metabolic functions to their host is also thought to explain why the remaining microsporidian core genes are shorter than their homologs in free-living fungi. An example can be seen in the sequenced genome of *Enc. cuniculi* where the average gene is 14 % shorter with respect to homologs in *S. cerevisiae* (Katinka et al. 2001; Slamovits, Fast, et al. 2004). It is thought that the absence of complex biochemical processes in the microsporidian cell eliminates the need for intricate protein-protein interactions. Thus the lack of selective pressure to maintain these protein-protein interacting epitopes culminates into their eventual deletion and consequent shortening of the coding DNA sequence. Ultimately, these shorter genes contribute to the overall genome shrinkage (Cavalier-Smith 2005).

1.4.3.3 Genome maintenance requirements

Cavalier-Smith (2005) postulated that the direction of microsporidian evolution towards a reduced genome could be driven by the host cell longevity. He argued that *E. bieneusi*'s reduced genome (~6 Mbp) may be partly due to the high turn-

over rate of intestinal mucosal cells it parasitizes. Thus, a smaller genome allows for faster replication to enable the production of mature spores within the limited lifespan of the host's intestinal mucosa cell (3-5 days) (Wright & Irwin 1982). Corroborating this hypothesis, *Edhazardia aedis*, the microsporidian with the largest genome currently known (~50 Mbp) parasitizes epithelial cells of the gastric caeca of adult mosquito midguts (Johnson et al. 1997). Interestingly, adult mosquitoes do not metamorphose and so their midgut epithelial cells have a relatively long lifespan (< 30 days) (Clements & Paterson 1981; Engel & Moran 2013)(Figure 1.6). Also, *Anncaliia algerae* has a large genome size of ~17 Mbp (Williams et al. 2008) and it has been recorded to infect human myocytes, a cell type with a long life-span (Millward et al. 1975). However the major host of *Anncaliia algerae* is the mosquito and infections in humans are considered to be accidental. (Monaghan et al. 2011; Watts & Chan 2014) (Figure 1.6). A Pearson correlation statistical analysis on known genome assembly sizes and host cell life span gives a value of -0.5, suggesting a negative correlation between genome assembly size and host cell longevity in accordance to Cavalier-Smith's hypothesis (Figure 1.6).

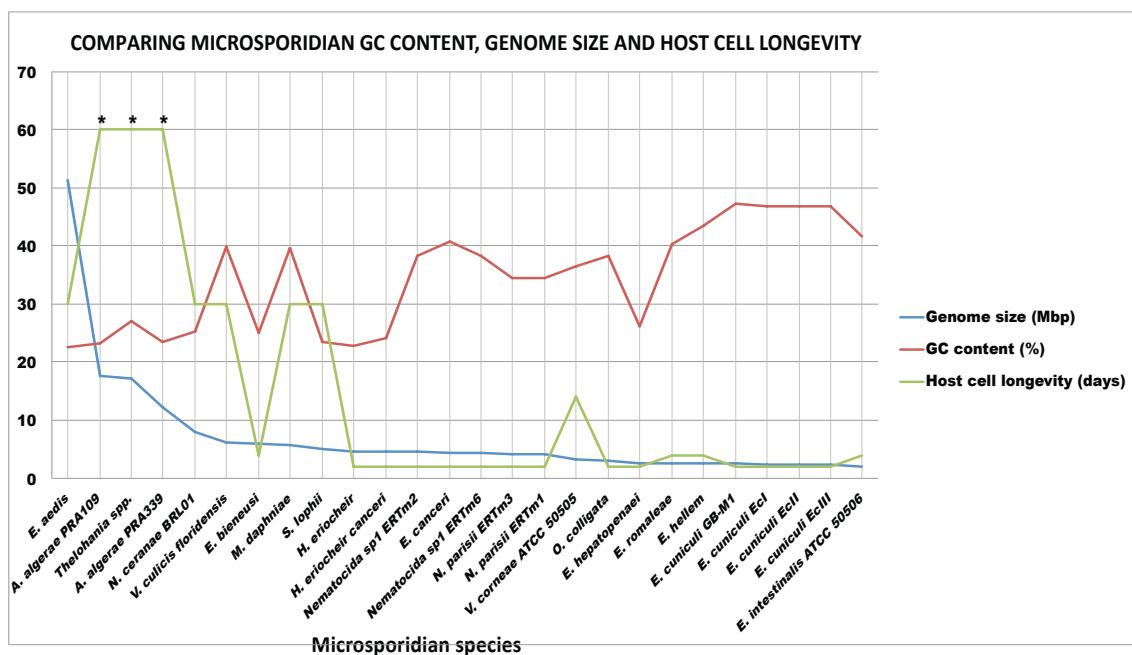


Figure 1.6: Mapping GC content and host cell longevity on genome size shows a positive correlation between genome size and host cell longevity and a negative correlation between GC content and genome size and host cell longevity. * Values are > 60 days. The host cell lifespans provided are that of healthy cells however this may be considerably lower in infected cells. Pearson correlation statistical analyses were performed here.

1.4.3.4 Effective population size decline and genetic drift:

Obligate endosymbiotic and intracellular parasitic lifestyles decreases the effective population size at every generation (Mendonça et al. 2011). This is because not all endosymbionts/intracellular parasites within a host are successfully transmitted to the next host. This creates a bottleneck in the effective population size between generations (Woolfit & Bromham 2003). This in turn increases the effect of weak evolutionary forces like genetic drift in fixating mutations within the new generation (Kelkar & Ochman 2012; Woolfit & Bromham 2003). These mutations could either be insertion or deletions giving rise to genome expansion or reduction respectively (Kelkar & Ochman 2012). In the case of Microsporidia, their intracellular parasitic lifestyle may have favoured genetic deletions. Here, genetic drift contributed in fixating deletion mutations and consequently reducing their genome size.

1.4.3.5 Loss of genetic redundancy

Free-living organisms are often exposed to unpredictable environmental factors that require them to respond appropriately in order to survive (Kitano 2007). The ability to develop mechanisms to survive in these variable environmental perturbations has been termed “robustness” by computational biologists (Kitano 2007). For single-celled organisms, robustness often exists on a molecular level which entails duplication of genes, protein domains and biochemical pathways (Mendonça et al. 2011; Papp et al. 2004). However, there is a reduced selective pressure to maintain the redundant duplicated gene. The paralogous proteins coded by these redundant genes are consequently riddled with mutations that can occasionally confer different isoelectric points or slightly variable active sites (Nowak et al. 1997).

Changes in the paralog's property can enable it to be activated in different environmental conditions or in the presence of certain substrates that may be unrecognised by the original paralogous protein. This new paralog could consequently rescue function in the event of a future deleterious mutation involving the original gene or change in environmental conditions such as pH or substrate scarcity and thereby conferring robustness (Papp et al. 2004; Mendonça et al. 2011). An example of this is seen in the three sugar-kinase paralogs in *Saccharomyces cerevisiae* that have been demonstrated to be activated in different environmental conditions and are even involved in pathways

not related to sugar metabolism [reviewed in (Kim & Dang 2005)]. There is however a minimal need for robustness for intracellular living organisms such as the Microsporidia as a result of protection from environmental factors provided by the host cell. The intracellular environment provides a comparatively predictable environment free from perturbations, leading to a reduced selective pressure to maintain redundant genes. The eventual deletion of these redundant paralogs contributes to the reduced genome size observed in intracellular organisms (Mendonça et al. 2011). In a recent comparative study looking at the genomes of *Enc. cuniculi* and *E. intestinalis*, Keeling and Corradi (2011) noted that the extremely reduced genome of *E. intestinalis* was predominantly due to the loss of gene paralogs in telomeric regions (Keeling & Corradi 2011).

1.4.3.6 Other contributors to the reduced genome in the Microsporidia

Other contributors to the reduced nature of the microsporidian genome include minimal or complete absence of introns and spliceosomal machinery genes. It has been reported that some microsporidian genomes that lack introns have some spliceosomal machinery genes still present. The proteins coded by these genes have been hypothesised to have other functions not related to splicing. Another characteristic of reduced microsporidian genomes is the absence of long intergenic regions (Desjardins et al. 2015).

1.5 The mitochondria and energy conservation

In extant model eukaryotes, the mitochondria have the crucial role of producing about 80 % of the cell's free ATP molecules (Berg et al. 2006). The citric acid cycle and oxidative phosphorylation are the mitochondrial biochemical processes that culminate in ATP production. Initially small carbon compounds such as pyruvate (C₃) (an end products of glycolysis) are channelled into the citric acid cycle. Prior to entering the citric acid cycle, which occurs within the mitochondrial matrix, pyruvate is oxidatively decarboxylated by pyruvate dehydrogenase into acetyl CoA (C₂) (Domingo et al. 1999). This reaction releases CO₂ as a biproduct and captures high-transfer-potential electrons in the form of NADH. Similar decarboxylation reactions ensue during the citric acid cycle which leads to the oxidation of acetyl CoA to two CO₂ molecules with the concomitant transfer of high transfer-potential-electrons to electron carriers, NAD⁺ and FAD. More specifically, each oxidized acetyl CoA compound, leads to the transfer of six and two electrons to three NAD⁺ and one FAD molecules respectively. Thus the citric

acid cycle generates a single FADH_2 and three NADH molecules (Krebs 1970; Barnes & Weitzman 1986; Berg et al. 2006).

During oxidative phosphorylation, NADH and FADH_2 , generated by the citric acid cycle are reoxidized by O_2 . Electrons released in the reoxidation of NADH and FADH_2 flow through membrane proteins situated in the inner mitochondrial membrane. These proteins are referred to the electron transport chain. These proteins together with their associated co-enzymes have special properties that take advantage of the double-membraned structure of the mitochondria to convert the electro-motive force of the electrons released by NADH and FADH_2 into proton-motive force. Thus, the flow of electrons through the electron transport chain is coupled with H^+ uptake from the mitochondrial matrix into the intermembrane space. The resulting unequal distribution of H^+ across the inner mitochondrial membrane creates a chemical and a charge gradient, which generates a proton-motive force (Yagi & Matsuno-Yagi 2003; Bianchi et al. 2004; Berg et al. 2006).

ATP synthase is another enzyme located on the inner mitochondrial membrane. It is made up of a H^+ -conducting part embedded into the mitochondrial membrane and a rotor part protruding into the matrix. The subunits responsible for catalysing the phosphorylation of ADP to form ATP are located in the rotor part of the enzyme. Interestingly, after the formation of ATP , this product continues to be tightly bound to the catalytic site of the enzyme. Due to the proton-motive force build-up across the inner mitochondrial membrane, H^+ ions in the intermembrane space flow from their region of high concentration back into the mitochondrial matrix, a region of low H^+ concentration through the H^+ -conducting part of the ATP synthase protein. The flow of H^+ through ATP synthase causes a conformational change of its catalytic subunits and the release of tightly bound ATP molecules. This, in turn frees the catalytic subunits for the binding of ADP and P_i to initiate the catalyses of another ATP molecule in a process known as rotational catalyses (Boyer 2000; Noji & Yoshida 2001). In summary, a single acetyl CoA compound can generate three NADH and one FADH_2 high energy molecules via the citric acid cycle. These molecules can in turn provide enough proton-motive force to power the formation of 9 ATP molecules via the electron transport chain and ATP synthase in a process known as oxidative phosphorylation (Berg et al. 2006).

1.6 Absence of the oxidative phosphorylation pathway in the Microsporidia

Evidence from independent microsporidian genome surveys has revealed that in the case where gene loss involves genes with known functions, these genes tend to have metabolic roles (Keeling et al. 2010; Keeling & Corradi 2011; Gill et al. 2008). Investigations by Cornman *et al.* (2009) demonstrated that there was a clear distinction in the number of metabolic genes between microsporidians and their free-living fungal counterparts. Results from this study suggested *Nosema* and *Encephalitozoon* genomes encode more structural and ribosomal proteins, and fewer metabolic enzymes in comparison to free-living fungal genomes. Many of these missing metabolic enzymes in microsporidians belong to the oxidative phosphorylation pathway, a metabolic process restricted to the mitochondria in extant model eukaryotes (Cornman et al. 2009).

Microsporidia, however, do not possess canonical mitochondria but rather a relic of what may be an ancestral mitochondria called the mitosome (Williams et al. 2002). Mitosomes lack oxidative phosphorylation capabilities but are thought to be involved in the generation of iron-sulphur clusters (Goldberg et al. 2008; Williams et al. 2008). With the absence of oxidative phosphorylation in Microsporidia, glucose is only partially metabolised to release about 20 % of its full ATP potential in a process known as glycolysis (Berg et al. 2006). This energy supply is complemented by the import of ATP molecules from the host with the help of ATP/ADP translocases (Nakjang et al. 2013; Hacker et al. 2014).

1.7 The role of microsporidian-host-mitochondria association in ATP acquisition

Association of intracellular parasites with the mitochondria of their hosts has been documented for both eukaryotic and prokaryotic pathogens (Horwitz 1983; Matsumoto et al. 1991; Tilney et al. 2001; Sinai et al. 1997). However this seems to be a trait employed only by certain species of parasitic lineages and not entire phyla. For example, whereas parasitophorous vesicles of *Toxoplasma gondii* are known to form close associations with the host mitochondria (Sinai et al. 1997), other apicomplexans such as *Plasmodium* do not form such associations (Bano et al. 2007). The bacterial pathogen *Chlamydia psittaci* is known to form close association with the host's mitochondria but members of the same genera such as *C. trachomatis* and *C. pneumoniae* have been shown not to form such

associations during infections (Horwitz 1983; Matsumoto et al. 1991; Tilney et al. 2001).

Similarly, microsporidian meronts belonging to various genera have been shown to form close associations with the host mitochondria. Examples include *Anisofilariata chironomi* (Tokarev et al. 2010), *E. bienewisi* (Canning & Hollister 1990) and *Encephalitozoon cuniculi* (Scanlon et al. 2004). In an extreme example of this type of interaction, not only did mitochondria of cells infected with *Thelohania solenopsae* aggregate around the developing meront but mitochondria in adjacent, non-infected cells also clustered at the periphery of the cell membrane closer to the infected cell (Sokolova et al. 2005). In most of the above-mentioned examples, it has been long hypothesised that the association between the developing parasite and the host mitochondria aids the parasite in harvesting ATP from the host. Studies by Hacker *et al.*, (2014) based on combinations of EM spatial quantification and mitochondria activity assays have hinted towards a protein-protein interaction between the parasitophorous vacuole of *Enc. cuniculi* and the host mitochondria. This study also revealed that the association between the parasite and host mitochondria increases the number of mitochondrial ATP export proteins, VDAC on the parasite-mitochondria interface thereby adding weight to the longstanding assumption that this association is to increase the microsporidian meront ATP sapping efficiency.

These results are indeed interesting, nonetheless there are still a number of unanswered questions regarding microsporidian energy conservation. The most obvious question is that despite the above-mentioned work by Hacker *et al.* (2014), the proteins involved in anchoring the microsporidian meronts to the host mitochondria are still unknown. Also, not all microsporidians associate with the host mitochondria thereby raising the question of how these lineages obtain energy from their host (Hacker et al. 2014). Arguably, the most intriguing question is whether the role of the mitochondrial association with the developing meront is solely for ATP exploitation or the mitochondria also plays a role in oxidizing NADH (a byproduct of glycolysis) released by the microsporidian cell.

1.8 The Enterocytozoonidae family

The Enterocytozoonidae are one of seven families of the Apansporoblastina suborder. Early development of the extrusion apparatus within a plasmodium

prior to sporogony is the major morphological characteristic that defines this microsporidian family as a distinct clade (Stentiford et al. 2011). Thus, in this family, the initial single meront undergoes multiple nuclear and internal organelle replications in the absence of cytokinesis thereby forming an enlarged plasmodium prior to sporogony (Vávra & Larson 1999). Another feature that seems to be exclusive to the Enterocytozoonidae family is the preference for an intranuclear infection site by some of its genera such as *Nucleospora*, *Enterospora*, *Enterocytozoon salmonis*, *Paranucleospora theridion* (see Table 1.1). In fact, out of nine well-studied species, only four are cytoplasmic (*Hepatospora* spp., *Enterocytozoon hepatopenaei*, *Desmozoon lepeophtherii* and *Enterocytozoon bieneusi*). *E. bieneusi*, however, forms tight associations with the host's nucleus. A second common characteristic, but not exclusive to the Enterocytozoonidae family, is their aquatic/marine-host preference. *E. bieneusi* is the only exception in this case even though there has been evidence to suggest an aquatic origin for it as well (Fournier et al. 2000) (Table 1.1).

Table 1.1: List of species within the Enterocytozoonidae family, site of intracellular infection, hosts and host's habitat and host's commercial value.

SPECIES WITHIN FAMILY ENTEROCYTOZONIDAE	HOSTS	TARGET HOST TISSUE/CELLS	INTRACELLULAR ENVIRONMENT	COMMERCIAL VALUE OF TARGET HOST (£/year)	REFERENCES	HOST'S HABITAT
<i>Desmozoon lepeophtherii</i> (= <i>Paranucleospora theridion?</i>)	Salmon louse (Atlantic salmon)	Epidermis and phagocytes	Cytoplasmic/ Intranuclear	837 million (In the UK only for the Atlantic salmon)	(Freeman & Sommerville 2009; Nylund et al. 2010; Seafish 2015)	Aquatic/Marine
<i>Enterocytozoon bieneusi</i>	Humans and wide range of mammalian and avian hosts.	Intestinal mucosa	Cytoplasmic	n/a	(Desportes et al. 1985)	Terrestrial
<i>Enterocytozoon hepatopenaei</i>	Black tiger shrimp Whiteleg shrimp	Hepatopancreatic cells	Cytoplasmic	2.46 billion (In Thailand only)	(Tourtip et al. 2009; Pongsri & Sukumasavin 2005)	Aquatic
<i>Nucleospora salmonis</i> (= <i>Enterocytozoon salmonis</i>)	Salmonids	Hemoblasts	Intranuclear	30 million (In Alaska only for the Chinook salmon)	(Hedrick et al. 1991; Chilmonczyk et al. 1991; ADFG 2016)	Aquatic/Marine
<i>Enterospora canceri</i>	European edible crab, hermit crab	Hepatopancreatic cells	Intranuclear	44 million (In the UK only for the edible crab)	(Stentiford et al. 2007; Stentiford & Bateman 2007)(Matthew et al. 2015)	Marine
<i>Enterospora nucleophila</i>	Gilt head sea bream	Enterocytes, inflammatory response cells	Intranuclear/ Cytoplasmic	30 million (In Spain only)	(Colloca & Cerasi 2005; Palenzuela et al. 2014; European Commission 2015)	Marine
<i>Hepatospora</i> spp.	European edible crabs, Chinese mitten crabs, pea crab	Hepatopancreatic cells	Cytoplasmic	7.5 billion (In China alone)	(Stentiford et al. 2011) (Weimin 2006)	Marine/Aquatic
<i>Nucleospora secunda</i>	<i>Nothobranchius rubripinnis</i>	Enterocytes	Intranuclear	Unknown	(Lom & Dykoá 2002)	Aquatic
<i>Nucleospora cyclopteri</i>	Lump fish	Hemoblasts	Intranuclear	20 million (In Iceland only for the caviar alone)	(Freeman et al. 2013) (FAO 2002)	Marine

1.8.1 Cytoplasmic-infecting species of the Enterocytozoonidae

Members of the Enterocytozoonidae family that reside in the host cytoplasm throughout their life cycle include *Enterocytozoon bieneusi* (Desportes et al. 1985), *Desmozoon lepeophtherii* (Freeman & Sommerville 2009), *Enterocytozoon hepatopenaei* (Tourtip et al. 2009), *Hepatospora eriocheir*

(Stentiford et al. 2011), *Hepatospora* sp. (Stentiford et al. 2011), all of which infect crustacean hosts (Table 1.1).

1.8.1.1 *Desmozoon lepeophtherii* (= *Paranucleospora theridion*)

Desmozoon lepeophtherii is a peculiar member of the Enterocytozoonidae because it defaults in the most definitive feature that unites members of this family, which is to mature from a merogonial plasmodium into a sporogonial plasmodium in the absence of cytokinesis. Instead, the merogonial plasmodium of *D. lepeophtherii* undergoes cytokinesis thereby forming individual uninucleate sporonts. The addition of this species to the Enterocytozoonidae was therefore mostly based on evidence from molecular analyses (Freeman et al. 2003; Freeman & Sommerville 2009). In its copepod host the salmon louse (*Lepeophtheirus salmonis*), *D. lepeophtherii* infects the innermost part of the epidermal layer beneath the cuticle and causes infected cells to hypertrophy and cluster into a bulbous structure called xenomas. This species is the only member of the Enterocytozoonidae currently known to form xenoma (Freeman & Sommerville 2009). Clinical symptoms of heavy infections include the production of unviable eggs in the female copepod host and opaque internal inclusions that are distributed throughout the host body (Freeman et al. 2003). Since the salmon louse is an ectoparasite of economically important salmon species and *D. lepeophtherii* infections are able to sterilize females, *D. lepeophtherii* has been suggested as a biological agent for use in alternative controls strategies for the salmon louse (Freeman & Sommerville 2011).

Systematics of *D. lepeophtherii* has not gone without controversy. A year after this parasite was described by Freeman and Sommerville (2009), a Norwegian group published a study describing a microsporidian parasite found in the same copepod hosts and parasitizing the same salmonid species and ascribed it a new genus and species name, *Paranucleospora theridion* (Nylund et al. 2010). Even though the authors recognised that the rDNA sequence of their newly named parasite had 99 % identity to that described in Freeman *et al.* (2003), they continued in assigning a new genus and species name. A reaction paper contesting this “second” taxonomic nomenclature and suggesting conspecificity between the parasites described by the two groups was published a year later (Freeman & Sommerville 2011).

As much as the microsporidian parasites described by the two groups share molecular and symptomatic similarities (similarity of rDNA and the formation of xenomas), the life cycle description presented by the two groups differ considerably. For instance, Nylund *et al.* (2010) described the parasite to have two life cycle stages alternating between the cytoplasm and nucleoplasm of a salmonid and copepod host (Nylund *et al.* 2010) whereas Freeman and Sommerville (2009) described a life cycle confined to the cytoplasm of the copepod host cell with no association with the host nuclei (Freeman & Sommerville 2009). The seeming synonymy of rDNA and symptomatic characteristics between the parasite described by the two groups and yet stark difference in the described life cycle stages demonstrates the peculiarity of this group of microsporidians and perhaps the need to employ stricter tools such as phylogenomics in resolving close evolutionary relationships. Chapter 3 of this manuscript is aimed at addressing this question using *Hepatospora* spp. as a case study.

1.8.1.2 *Hepatospora* spp.

A study of the causative agent of the tremor disease in native Chinese mitten crabs (*Eriocheir sinensis*) led to the discovery of a microsporidian parasite infecting the hepatopancreas of these crabs (Wang & Gu 2002). Despite this parasite not being directly associated with tremor disease, it was considered to be equally detrimental to the animal's health due to the sloughing of the hepatopancreatic tissue observed in infected animals (Wang & Chen 2007). In their study, Wang and Chen (2007) were unsuccessful in amplifying the parasite's rDNA sequences for phylogenetic analysis, which led them to base their taxonomic classification entirely on ultrastructural features. As such, the mitten crab parasite was assigned to the *Endoreticulatus* genus, (*Endoreticulatus eriocheir*) but the authors mentioned the need to confirm their findings with molecular data in the future.

In 2011 an environmental survey aimed at identifying the parasitic profile of invasive Chinese mitten crabs in the UK identified a microsporidian with identical morphological characteristics as the one described by Wang and Chen (2007). Phylogenetic studies performed with rDNA however showed this parasite to cluster with members of the Enterocytozoonidae family rather than *Endoreticulatus* (Stentiford *et al.* 2011). In their study Stentiford *et al.* (2011)

observed that the ultrastructural and developmental features characterized by this parasite were more reminiscent of the Enterocytozoonidae family. Some of these features included smaller spore size, and the early development of the extrusion apparatus in a plasmodium consisting of several sporonts maturing in synchrony (Stentiford et al. 2011). Also, the latest authoritative study looking at systematics within the microsporidian phylum placed *Endoreticulatus* within the terrestrial microsporidian clade, hence it is unlikely that a parasite infecting an aquatic host (90 % of mitten crabs in China), would be a member of this genus (Vossbrinck & Debrunner-Vossbrinck 2005; Wang & Chen 2007). In light of this ultrastructural and molecular evidence, Stentiford *et al.* (2011) updated Wang and Chan's description of *Endoreticulatus eriocheir* and erected a new genus, *Hepatospora* within the Enterocytozoonidae family. Whereas this genus was meant to encompass cytoplasmic Microsporidia infecting the hepatopancreas of decapod hosts the authors suggested that the erection of a higher taxonomic rank (e.g. Family Hepatosporidae) would be most appropriate but refrained from doing so due to the present confusion at higher taxonomic levels within the microsporidian phylum (Vossbrinck & Debrunner-Vossbrinck 2005).

Subsequent disease profile surveys performed on UK crab populations also revealed *Hepatospora*-like microsporidians parasitizing the hepatopancreas of the pea crab (*Pinnotheres pisum*) and the edible crab (*Cancer pagurus*) (Bateman et al. 2011; Longshaw et al. 2012). The particular microsporidian species observed in edible crabs had already been mentioned in Stentiford *et al.* (2011) and been assigned the name *Hepatospora* sp. Infections in both cases portrayed a similar hyperparasitism of the cytoplasm of the host's epithelial cells, which in turn led to the occasional degeneration of the hepatopancreatic tubule as previously observed in *H. eriocheir*. In all cases, the infected crabs were asymptomatic (Bateman et al. 2011; Longshaw et al. 2012). Interestingly, despite the congruence in pathology displayed by these parasites in all three host species and 100 % rDNA similarity (Bateman et al. 2011), these microsporidians differed in polar tubule arrangement and karyotypic traits. Since these traits have been pivotal in taxonomic name assignment within the phylum Microsporidia (Vávra 1976; Sprague 1977; Larsson 1986; Sprague et al. 1992)(Section 1.3.3.), the putative polymorphism displayed by these crab microsporidians posed serious questions to systematics within this phylum. Chapter 4 of this thesis uses a

phylogenomic approach to tackle this question and dissect the issue of morphological plasticity within the phylum Microsporidia.

1.8.1.3 *Enterocytozoon hepatopenaei*

Enterocytozoon hepatopenaei is arguably the highest profile member of this family due to its association with Slow Growth Syndrome (SGS), which is causing serious financial losses in the Thai fishing industry within the region of \$300 million per year (Chayaburakul et al. 2004). Since its emergence in 2001, SGS in farmed black tiger shrimps, *Penaeus monodon* had been linked to a plethora of etiological agents including viruses, bacteria and microsporidia but it was only in 2009 that Tourtip *et al.* (2009) provided a full description for the associated microsporidian agent. However this study did not include any statistical evidence to associate *E. hepatopenaei* infections directly to SGS (Tourtip et al. 2009). It has been suggested that *E. hepatopenaei* infections in black tiger shrimps may be opportunistic and possibly exacerbate an already established SGS caused by bacterial and viral agents (Chayaburakul et al. 2004). This SGS surge in native Thai black tiger shrimp populations led farmers to switch to the whiteleg shrimp *Penaeus vannamei*.

Interestingly, not long after the introduction of the whiteleg shrimp in Thai fisheries, they were also observed to exhibit White Faeces Syndrome (WFS), which was often accompanied by *E. hepatopenaei* infections (Tangprasittipap et al. 2013). The authors, however, noted that even though *E. hepatopenaei* is not the causative agent of WFS, high parasitemia in the whiteleg shrimp would burden the animal with increased energy demand and therefore lead to stunted growth (Tangprasittipap et al. 2013). In both the black tiger shrimp and whiteleg shrimp hosts, *E. hepatopenaei* infections occur in the cytoplasm of all cell types of the hepatopancreatic tubule epithelium (embryonic, blister, reserve and fibrillar cells) (Tourtip et al. 2009; Tangprasittipap et al. 2013). However, spore formation in the whiteleg shrimp seems to be exclusive to blister cells (Tangprasittipap et al. 2013). High infection levels observed in the recently introduced whiteleg shrimp has hinted to the presence of a natural reservoir species and current efforts are aimed at developing robust molecular markers for the early detection of *E. hepatopenaei* in ponds but also to identify the unknown natural reservoir(s) and eliminating it (them) from fish farming ponds (Tangprasittipap et al. 2013). To this end, Chapter 2 of this thesis will describe the work undertaken to sequence

and annotate the genomic DNA of *E. hepatopenaei* (and the genomes of other Enterocytozoonidae), which is currently being used by collaborators in developing sensitive and specific primers for the early detection of this microsporidian in Thai fishing ponds.

1.8.1.4 *Enterocytozoon bieneusi*

Enterocytozoon bieneusi was the first species to be added to the Enterocytozoonidae family (Sprague et al. 1992) and has received particular attention since its discovery in 1985 (Desportes et al. 1985). It is thought to contain at least 64 human-infecting genotypes (Feng et al. 2011; Matos et al. 2012; Fayer & Santin-Duran 2014), 34 of which lack host specificity and thus can infect both humans (Tumwine et al. 2005; Wichro et al. 2005; Chacin-Bonilla et al. 2006; Leelayoova et al. 2005) and a broad range of other vertebrate hosts including wild and domesticated animals (del Aguila et al. 1999; Lores et al. 2002; Haro et al. 2006; Santín & Fayer 2011). The promiscuity of *E. bieneusi*'s host selection has raised public health concerns and led to several investigations into its zoonotic transmission (Sulaiman et al. 2003; Sulaiman et al. 2004; Widmer & Akiyoshi 2010; Henriques-Gil et al. 2010). These studies were crucial in highlighting the tremendous genotype diversity displayed by *E. bieneusi* and distinguishing between host-specific and non-host specific genotypes, with host-specific and promiscuous genotypes frequently observed in developed countries and underdeveloped countries respectively (Dengjel et al. 2001; Breton et al. 2007; Stark et al. 2009; Mori et al. 2013).

E. bieneusi is the most prevalent human-infecting microsporidian (Mathis et al. 2005) and is known to infect epithelial cells of the upper gastrointestinal tract and cause chronic diarrhoea in HIV and organ transplant patients (Sax et al. 1995; Rabodonirina et al. 1996; Kelkar et al. 1997; Guerard et al. 1999; Goetz et al. 2001; George et al. 2012) which has been reported to persist for more than two years in some cases (Weber et al. 1992). Although the infection is limited to epithelial cells, it is not confined to just the gastrointestinal tract as some infections have been reported in lungs of bone marrow transplant patients (Kelkar et al. 1997). Chronic infections can cause the fusion and blunting of gastrointestinal villi that consequently leads to malabsorption and wasting syndrome typically seen in advanced cases of HIV. In healthy adults, *E. bieneusi*

infections are characterised by self-limiting diarrhoea (López-Vélez et al. 1999; Müller et al. 2001; Wichro et al. 2005).

The nested phylogenetic position of *E. bienersi* within the *Hepatospora/Nucleospora/Desmozoon/Enterospora* clade has raised questions about the possible role of an aquatic host in its transmission (Tourtip et al. 2009; Freeman & Sommerville 2009; Nylund et al. 2010; Stentiford et al. 2011; Stentiford et al. 2013; Freeman et al. 2013). Evidence in support of the aquatic origin of *E. bienersi* hypothesis include a previous study that detected *E. bienersi* in water samples collected from the French Seine river (Fournier et al. 2000). However, it must be stressed that this study did not provide any microscopy-based evidence for the presence of microsporidian spores and molecular probes revealed minimal contamination (Fournier et al. 2000). A similar survey performed on zebra mussels (*Dreissena polymorpha*) sampled from the Irish Shannon River also revealed low levels of *E. bienersi* spores in collected samples without evidence of active infection of mussels (Graczyk et al. 2004). The *E. bienersi* spores detected by Fournier *et al.* (2000) and Graczyk *et al.* (2004) could have however come from discharged domestic water or animal sources and not necessarily from an aquatic host.

A survey of *E. bienersi*'s genomic DNA failed to identify most genes involved in the glycolytic, trehalose and pentose phosphate pathways and fatty acid synthesis, which led the authors to suggest a sole reliance of *E. bienersi* on its host for its energy supply via ATP transporters (Akiyoshi et al. 2009). The absence of genes encoding proteins involved in these core metabolic pathways was indeed curious and warranted the resequencing of the *E. bienersi* genome (Keeling et al. 2010) which produced similar results. Even if this microsporidian acquires energy from its host during its merogonial stage, absence of the two key energy-conservation pathways (pentose phosphate pathways and glycolysis) poses the question of how it obtains energy for germination, an energy intensive process requiring ATP to drive polar tube polymerisation (Weidner & Byrd 1982; Weidner et al. 1995). This discovery has profound significance for our view of the minimal requirements for a eukaryotic cell and makes this family a crucial group for the purposes of eukaryotic reductive evolutionary studies. Chapter 3 of this thesis will compare publicly available genomic data from this parasite with genomic data produced during this PhD of closely related species to investigate

whether this unusual loss of glycolysis is unique to *E. bieneusi* or a common trait within the Enterocytozoonidae.

1.8.2 Intranuclear species of the Enterocytozoonidae

Association with specific host organelles is a common trait of microsporidians with the majority of associations being formed with the host mitochondria (Shadduck & Pakes 1971; Larsson 1980; Canning & Hollister 1992; Scanlon et al. 2004). This has been shown to aid the developing meront/sporont to scavenge ATP from the host cell (Hacker et al. 2014). The second most common association is with the host nuclei which is predominantly accompanied by nuclear hypertrophy [*Spraguea lophii* (= *Nosema lophii*) (Weissenberg 1976)], karyokinesis [*Glugea anomala* (Canning & Hazard 1982), *Thelohania bracteata* (Liu 1972)] and increase in chromosomal volume [*Thelohania* and *Octospora* sp. (Pavan et al. 1969)]. This association is often limited to the juxtaposition of the developing meront to the host nucleus and in some cases the development of the microsporidian cell within a nuclear invagination but still retaining some contact with the host's cytoplasm [*Chytridiopsis socius* (Sprague et al. 1972), *Enterocytozoon bieneusi* (Desportes-Livage et al. 1991)]. The first microsporidian ever to be identified to inhabit the nucleus of its host was *Microsporidium rhabdophilia* (Modin, 1981) which parasitized rodlet cells of several salmonid species: *Oncorhynchus tshawytscha*, *O. kisutch*, *O. mykiss gairdnerii* and *O. mykiss irideus* (Modin 1981). Complete intranuclear localization had been reported for some *Nosema* species but these were thought to be accidental infections rather the norm [original French publications (Takizawa et al. 1973; Loubès & Akbarieh 1978) and reviewed in English in (Vávra & Larson 2014)]. Association with the host nucleus is also common among the major protozoan group, Alveolata. More specifically, certain member species of the *Amoebophyra*, *Toxoplasma*, *Plasmodium* and *Eimeria* genera, have been repeatedly observed to inhabit the nucleus of their hosts (Davis et al. 1957; Atkinson & Ayala 1987; Barbosa et al. 2005; Bano et al. 2007; Bould et al. 2009; Miller et al. 2012). Below are descriptions of current intranuclear members of the Enterocytozoonidae family.

1.8.2.1 *Nucleospora* spp.

Nucleospora salmonis was originally described by Hedrick *et al.* (1991) and redescribed in the same year by Chilmonczyk *et al.* (1991). In their study, Chilmonczyk *et al.* (1991) assigned this microsporidian to the *Enterocytozoon* genus but a detailed comparative study between the developmental cycle of the parasite and *E. bieneusi* showed that there was too little similarity between the two species for them to be grouped into the same genus (Desportes-Livage *et al.* 1996). A comparative study of rDNA fragments between *E. bieneusi* and *N. salmonis* also corroborated the findings of Desportes-Livage *et al.* (1996) and advocated the retention of the new genus name, *Nucleospora* (Docker *et al.* 1997). *N. salmonis* infects lymphoblasts and plasmoblasts of the Chinook salmon (*Oncorhynchus tshawytscha*), silver trout (*Oncorhynchus nerka*), steelhead trout (*Oncorhynchus mykiss*), lake trout (*Salvelinus namaycush*), brook trout (*Salvelinus fontinalis*) and Atlantic salmon (*Salmo salar*) which in turn causes hypertrophy of the large intestine, spleen and kidney, a condition known as plasmacytoid leukemia (Morrison *et al.* 1990; Hedrick *et al.* 1990; Mullins *et al.* 1994; Bravo 1996; Gresoviac *et al.* 2000; Foltz *et al.* 2009; Badil *et al.* 2011; Freeman *et al.* 2013). In Chinook salmon, *Nucleospora salmonis* infections are common in juvenile fish and has been associated with severe acute anaemia (Elston *et al.* 1987; Hedrick *et al.* 1990; Hedrick *et al.* 1991).

Similarly, *Nucleospora cyclopteri* (Freeman *et al.*, 2013) causes renomagalay in the Icelandic lumpfish (*Cyclopterus lumpus*) and is associated with exophthalmos (swelling and protrusion of the eyeball) (Freeman *et al.* 2013). The extent of the impact of *Nucleospora cyclopteri* infections on Icelandic lumpfish fisheries will need a more detailed survey in the future but current evidence suggests a widespread geographical distribution of the *Nucleospora* as occurrences span from the North Atlantic to the South Pacific Ocean (Bravo 1996; Freeman *et al.* 2013) prompting some authors to suggest that all salmonid and perhaps some non-salmonid fish species may be susceptible to *Nucleospora* infections (Gresoviac *et al.* 2000). Interestingly, the second species assigned to this genus was isolated from ornamental fish, *Nothobranchius rubripinnis*, originating from Tanzania (Lom & Dykoá 2002). Since *Nothobranchius rubripinnis* is a warm water African fish with no possible connections to lumpfishes, trouts and salmonids, *Nucleospora secunda* infections in this fish species further corroborate suggestions of a geographical widespread distribution for this genus (Gresoviac

et al. 2000; Khattra et al. 2000). Unlike *N. salmonis* and *N. cyclopteri* that parasitize hemoblasts, *N. secunda* infects erythrocytes of its host (Lom & Dykoá 2002). In the absence of molecular data and information about host pathology, the taxonomic placement of *Nucleospora secunda* was entirely based on its ultrastructural characteristics which limits parallels that could be drawn between this species and other members of this family (Lom & Dykoá 2002).

Phylogenetic analysis based on rDNA of *Nucleospora* species revealed that members of this genus branched as a monophyletic group and as a sister group to the *Desmozoon* clade, whose member species are thought to possess a life cycle that alternates between a parasitic copepod and a fish host (Freeman & Sommerville 2009; Nylund et al. 2010). In his discussion, Freeman *et al.* (2013) hinted at the possibility of a similar undescribed life cycle for *Nucleospora cyclopteri* and the likelihood for a similar copepod to serve as a reservoir. In summary, there is evidence to support the notion that fish disease as a consequence of infections by members of the *Nucleospora* genus has become increasingly common in recent years thereby making them extremely important pathogens of both farmed and wild commercial fisheries and perhaps for the ornamental fish industry as well (Hedrick et al. 1991; Higgins et al. 1998; Lom & Dykoá 2002; Stentiford et al. 2013; Kent et al. 2014).

1.8.2.2 *Enterospora nucleophila*

In 2014 a Spanish research group investigating an emaciative syndrome causing mortality in as many as 1 % of farmed juvenile gilthead sea bream (*Sparus aurata*) per day isolated a microsporidian parasite as the putative aetiological agent (Palenzuela et al. 2014). The parasite, described as *Enterospora nucleophila* parasitized the cytoplasm and nuclei (albeit occasionally) of enterocytes and macrophages but had a particular affinity for the host's rodlet inflammatory response cells (Reite 2005; Palenzuela et al. 2014). In their study Palenzuela *et al.* (2014) revealed that *E. nucleophila* infections, which were accompanied by stunted growth and host mortality, were particularly prominent in fish sampled during the cold winter months (Palenzuela et al. 2014). Cold temperatures sustained by farmed gilthead sea bream in winter months had been previously linked to reduced fish immunity and a concomitant increased susceptibility to bacterial and viral infections that cause a plethora of symptoms

collectively known as Winter Disease Syndrome (WDS) (Doimi 1996; Tort, Rotllant, et al. 1998; Domenech et al. 1999). This led the authors to suggest that *E. nucleophila* infections may be opportunistic and arise as a consequence of Winter Disease Syndrome. Interestingly, all histopathological symptoms of *E. nucleophila* infections evidenced by Palenzuela *et al.* (2014) such as hyperproliferation of intestinal mucosal cells, hypertrophy of the lamina propria and accumulation of inflammatory cells to the intestinal submucosa have previously been documented to also be symptoms associated with Winter Disease Syndrome (Tort, Rotllant, et al. 1998; Tort, Padros, et al. 1998). As such, one cannot help but to wonder if the emaciative syndrome described in Palenzuela *et al.* (2014) is not in fact a juvenile fish version of WDS and that *E. nucleophila* is not opportunistic *per se* but another aetiological agent of WDS. It is difficult to arrive at a definitive conclusion here since key physical symptoms of the WDS such as upside down swimming of fish were not described in Palenzuela *et al.* (2014). This is however understandable as authors had received most samples for their study pre-necropsied by farm staff from different locations (Palenzuela et al. 2014) and hence may not have been able to observe the fish prior to necropsy.

Taxonomic assignment of *E. nucleophila* to the Enterocytozoonidae family was supported by its ultrastructural development, which comprised of early development of polar filament in a sporogonial syncytium in direct contact with the host nucleoplasm or cytoplasm. Spore features characteristic of the *Enterospora* include small spore size of 1 x 1 µm and 5-6 polar filament coils. Just as some of the other species in this family such as *Desmozoon cyclopteri*, *E. nucleophila* was observed to exhibit both an intranuclear and a cytoplasmic life cycle (Nylund et al. 2010; Palenzuela et al. 2014). However, whereas the intranuclear and cytoplasmic life cycles of *D. cyclopteri* occur in a teleost and a copepod host respectively, both life cycles occur in the teleost host in the case of *E. nucleophila* (Nylund et al. 2010; Palenzuela et al. 2014). Also, since *Enterospora canceri*, the only other species within the *Enterospora* genus, is a crab parasite and crustaceans are routinely used as live prey in the early stage of marine fish farming, the involvement of a secondary crustacean host in the transmission of *E. nucleophila* cannot be ruled out (Stentiford & Bateman 2007; Stentiford et al. 2007; Palenzuela et al. 2014).

Phylogenetic analysis performed with rDNA sequences of *E. nucleophila* also strongly supported its placement within the Enterocytozoonidae family but it did not form a monophyletic clade with the other member of the *Enterospora* genus, *Ent. canceri* (Palenzuela et al. 2014). Just as in the case of *Enterocytozoon hepatopenaei*, the authors' decision to add the new species to an already described genus was a conservative approach in light of the current confusion of higher taxonomic ranks within the phylum Microsporidia (Vossbrinck & Debrunner-Vossbrinck 2005; Tourtip et al. 2009; Palenzuela et al. 2014).

The value of gilthead sea bream market in Spain is around €40 million with the entire European market value around €363 million (European Commission 2015) (Table 1.1). The economic importance of this fish for Spain and Europe as whole and the potential threat to yield by *E. nucleophila* advocates for further studies of this microsporidian parasite to resolve its life cycle (in particular to find the intermediate crustacean host) but also to create robust principles for taxonomic assignments within the Enterocytozoonidae, which could later feed into legislation.

1.8.2.3 *Enterospora canceri*

This microsporidian was described as the first ever known intranuclear species to parasitize the nucleus of an invertebrate host and still remains the only species with such a target host and intracellular compartment to date (Stentiford et al. 2007; Stentiford & Bateman 2007). *Enterospora canceri* has so far been isolated from the hepatopancreas of the edible (*Cancer pagurus*) and hermit crab (Stentiford et al. 2007; Stentiford & Bateman 2007). Among these, only the edible crab is commercially exploited with the UK being the largest exporter of the edible crab in Europe. It is estimated that £44 million worth of edible crabs were landed in UK waters in 2014 alone (www.gov.uk/government/statistical-data-sets). Unlike the edible crab, the hermit crab is not exploited as a direct food source for human consumption but plays an important role in the marine/estuarine food chain as a benthic organism and may be preyed upon by the edible crab (Lawton 1989; Ramsay et al. 1997). The hermit crab is an important species for understanding animal personality, behavioural plasticity and for modelling evolutionary and ecological traits (Vannini & Cannicci 1995; Briffa & Elwood 2005; Berke et al. 2006; Elwood & Appel 2009; Briffa et al. 2013). Due to their sedentary behaviour and their ability to filter feed, hermit crabs are also used as

sentinel species or bioindicators for accessing the geographical distribution of environmental pollutants in coastal areas (Gerlach et al. 1976; Boon et al. 2002; Law et al. 2003; Morris et al. 2004; Tian et al. 2010; Sant'Anna et al. 2014).

In both crab species *Ent. canceri* infections were asymptomatic with a maximum prevalence of 3.45 % (Stentiford et al. 2007; Stentiford & Bateman 2007). However it must be mentioned that the edible crabs used in this study were destined for the commercial market and hence were all above the minimal landing carapace size of 140 mm. As such, juvenile crabs were automatically excluded from this study (Stentiford et al. 2007). A subsequent survey comparing disease profiles between edible crabs above and below the landing size revealed that prevalence of *Ent. canceri* infections could increase to 6-7 % in both subpopulations (Bateman et al. 2011). However, the exact impact on edible crab yield of the seasonal elevated prevalence of *Ent. canceri* infection could not be assessed (Bateman et al. 2011). Nonetheless, the single heavy infection accompanied by degeneration of the hepatopancreatic tissue observed in Stentiford and Bateman (2007) suggests that despite the absence of physical symptoms, *Ent. canceri* infected crabs are moribund.

The placement of *Ent. canceri* within the Enterocytozoonidae family was supported by ultrastructural evidence that revealed this parasite to form a sporogonial plasmodium with the early appearance of extrusion apparatus precursors (Stentiford et al. 2007; Stentiford & Bateman 2007). Phylogenetic studies based on rDNA sequences so far have repeatedly positioned *Ent. canceri* as the closest relative of the human infecting microsporidian, *E. bienersi* (Stentiford et al. 2011; Freeman et al. 2013; Stentiford et al. 2013; Palenzuela et al. 2014). In their discussion, Palenzuela *et al.* (2014) described *Ent. canceri* to possess both a cytoplasmic and intranuclear life stage. It must however be clarified that in the original description of this species all life stages were observed within the host's nucleoplasm and the only occasions spores were observed in the cytoplasm was when the host's nucleus was overfilled with the parasite's spores leading to the fracturing of the nuclear membrane and an overspill of spores into the host's cytoplasm. An overfilled host cell eventually ruptures and releases infective spores into the tubule lumen (Stentiford et al. 2007; Stentiford & Bateman 2007). The only difference between *Ent. canceri* samples collected from the two crab species was that precursors of the polar tubule were observed to surround the parasite's vacuole during sporoblast development of the hermit

crab parasite but no such observation was made for the edible crab parasite (Stentiford et al. 2007; Stentiford & Bateman 2007).

In the absence of oxidative phosphorylation pathways, microsporidians are considered to rely on intrinsic glycolysis and a plethora of transporters including horizontally acquired ATP transporters for their energy needs (Katinka et al. 2001; Williams et al. 2002; Tsaousis et al. 2008)(Section 1.6). Consistent with this notion, many microsporidian species have been observed to form intimate physical associations with the host mitochondria, the principal source of ATP in the host cell and this association has been demonstrated to increase the parasite's energy sapping efficiency (Hollister et al. 1996; Scanlon et al. 2004; Sokolova et al. 2005; Tokarev et al. 2010; Hacker et al. 2014). The development of *Ent. canceri* within the host nucleus (a compartment physically walled off from the host mitochondria) and its close phylogenetic affinity with *E. bienersi*, a species whose genome was recently found to be devoid of genes coding for enzymes involved in core metabolic processes such as glycolysis, pentose phosphate and trehalose pathways, makes the case of *Ent. canceri* a curious one indeed (Akiyoshi et al. 2009; Keeling et al. 2010). It raises the question of whether *Ent. canceri* has lost glycolytic capabilities like its closest relative, *E. bienersi*, and if so poses the question of how this parasite obtains energy from its intranuclear subcellular environment. Chapter 4 of this manuscript will try to resolve these questions by employing phylogenomic and molecular cloning techniques.

1.9 The nucleus as a niche

The nucleus is a double membraned subcellular compartment rich in nucleic acids in the form of DNA and RNA. There is evidence to suggest that some intranuclear parasites use host nucleic acids as a source of sugar, nitrogen and phosphates. For example, intranuclear infections of bacterial parasite, *Candidatus Nucleicultrix amoebiphila*, *Candidatus Endonucleobacter bathymodioli* and some *Holospira* spp. are accompanied by the disappearance of the host's heterochromatin (Görtz 1986; Zielinski et al. 2009; Schulz et al. 2014). The phagosome of some intranuclear *Amoebophyra* spp. have also been observed to contain host chromatin (Miller et al. 2012). Past studies on subcellular concentrations of sugars, ions, lipids, ATP and proteins have hinted to high levels of these substances in the nucleus (Naora et al. 1962; Eichberg et

al. 1964; Miller & Horowitz 1986; Cascianelli et al. 2008). It is therefore understandable that some parasites have evolved to inhabit this subcellular compartment. It must however be mentioned that subcellular concentration of these substances vary between cell types and between species. As these studies were not performed on the host species infected by intranuclear microsporidia nor the cell types they parasitize, it cannot be conclusively established if nutrient acquisition is a reason behind the evolution of intranuclear living in the Microsporidia. This however seems to be the case for many intranuclear bacterial parasites (Görtz 1986; Zielinski et al. 2009; Schulz et al. 2014). Apart from serving as a rich source of nutrients, an intranuclear niche provides refuge for the invading parasite from cytoplasmic harmful substances that cannot be actively transported via the nuclear pores or diffuse across the nuclear envelope (Bonner 1975; Pongponratn et al. 1998). Cytoplasmic infections are often cleared by the cell via an innate immune mechanism known as autophagy. Here, an invading parasite is transported to the lysosome and destroyed (Huang & Brumell 2014). As the nucleus is devoid of such defence mechanisms, it is possible that it provides an intranuclear parasite with a perturbation-free niche ideal for replication (Schulz & Horn 2015). Another potential benefit of intranuclear living is that it provides the parasite with proximity to the host's genetic material which the parasite can manipulate to program the host cell to favour its survival (Hori et al. 2008). Despite the obvious benefits of intranuclear inhabitation, it is still odd that some Microsporidia have evolved to live inside the host nucleus, a subcellular compartment physically walled off from the host mitochondria. This is because the developing meronts of members of this phylum are often found in close association with the host's mitochondria. There is current evidence to suggest that this association with the host's mitochondria (The main ATP producing sites in the cell) increases the microsporidia's ATP sapping efficiency (Hacker et al. 2014).

1.10 Importance of studying the Enterocytozoonidae family

1.10.1 Potential for zoonotic transmission in humans

The occurrence of the same *E. bienersi* strains in both animals and humans have been repeatedly reported but a life cycle involving both a human and animal hosts has yet to be elucidated (Dengjel et al. 2001; Reetz et al. 2002; Sulaiman et al. 2004; Haro et al. 2005; Mori et al. 2013; Zhao et al. 2015). Some authors have

suggested and in some cases confirmed foodborne, waterborne and even airborne transmission modes. However, the repeated nesting of *E. bienersi* within a crustacean- and fish-infecting microsporidian clade (Enterocytozoonidae clade) on rDNA based phylogenetic analysis is suggestive of a zoonotic transmission involving an aquatic reservoir host (Haro et al. 2005; Slodkowitz-Kowalska et al. 2006; Tourtip et al. 2009; Freeman & Sommerville 2009; Nylund et al. 2010; Stentiford et al. 2011; Decraene et al. 2012; Stentiford et al. 2013; Freeman et al. 2013). Moreover, alternation between vertebrate and invertebrate hosts has been observed for some species within the Enterocytozoonidae clade, corroborating the likelihood of an aquatic invertebrate reservoir host involved in the transmission of *E. bienersi* (Freeman & Sommerville 2009; Nylund et al. 2010). Considering the mortality attributed to *E. bienersi* infections especially in immunocompromised patients and the increasing reports of such infections in immunocompetent individuals (López-Vélez et al. 1999; Sewankambo et al. 2000; Müller et al. 2001; Wichro et al. 2005), identification of its putative aquatic reservoir host and a better understanding of its transmission mode would be invaluable for management of the disease and could perhaps elucidate on an animal model for its artificial propagation as current efforts to culture this parasite in a laboratory setting have proven futile.

1.10.2 Diseases of commercially important fisheries

Table 1.1 shows the economic importance of fisheries affected by infections by members of the Enterocytozoonidae family, which in turn highlights the need to safeguard these fisheries against diseases in order to sustain and maximize profitability for the communities that depend on them for their livelihood and for the countries involved as a whole. To achieve this, governments are responsible for legislating policies to protect local uninfected but susceptible fish populations from being contaminated with parasites carried by a foreign fish population. An example of such legislation in the UK is the “Import of Live Fish Act 1980” which requires special licenses and quarantine measures for the import of live non-native fisheries. Extremes of these safeguarding measures could also be in the form of trading embargos on farms or countries whose fisheries have been identified to harbour an infectious parasite or the destruction of entire fish farm yields confirmed to harbour an infectious disease as stipulated by the European Commission Council Directive 93/53/EEC for the control of spring viraemia of

carp. These safeguarding measures can lead to massive revenue losses for farms and countries involved and so such policies ought to be informed by sound science. In line with this argument, it is important that the taxonomy of infectious parasites of these economically important fisheries are resolved to the strain level as taxonomic names of parasites are fed into these safeguarding policies. For instance, a taxonomic name that erroneously describes closely related non-pathogenic and pathogenic strains as a single entity could lead to unfair trading bans on farms or countries whose fisheries harbour non-pathogenic strains. As such, proper systematics of pathogenic parasites is crucial to prevent unfair trading restrictions (Stentiford et al. 2014). Systematics within the phylum Microsporidia is under current reform but this reform should be perhaps prioritised for the Enterocytozoonidae family as each of its member species is an aetiological agent in at least one multimillion-pound fishery or a major cause of mortality in immunocompromised human patients (Table 1.1). Moreover, the similarity in morphology and their elusive life cycles that are characterised by dissimilar developmental features in different host species makes member species of the Enterocytozoonidae family even more difficult to distinguish from each other for taxonomic name assignment (Stentiford et al. 2007; Stentiford & Bateman 2007; Freeman & Sommerville 2009; Nylund et al. 2010; Freeman & Sommerville 2011). Molecular approaches based on rDNA sequences have been pivotal in the recent overhaul of microsporidian systematics, however rDNA sequences have proven not to be useful in differentiating between closely related species (Vossbrinck & Debrunner-Vossbrinck 2005; Stentiford et al. 2013). With the emergence of whole genome data for members of the phylum Microsporidia efforts are being directed in identifying molecular markers other than rDNA and using phylogenomics for the assessment of relatedness between species (Cuomo et al. 2012; Pan et al. 2013; Nakjang et al. 2013; Haag et al. 2014). In a similar effort, Chapter 2 of this thesis describes the sequencing of the genomes of 3 species of the Enterocytozoonidae. Furthermore, chapter 2 and 4 focus on the use of a phylogenomic approach in distinguishing between members of the Enterocytozoonidae family.

1.10.3 Potential biological control agents

Desmozoon lepeophtherii was included in a UK patent issued in 2002 for the biological control of sea lice: "Microbiological control of sea lice" United Kingdom

patent GB2371053, international patent PCT/GB02/00134. At the time, *D. lepeophtherii* was known to cause the production of sterile eggs in the sea louse (*L. salmonis*), which parasitized farmed Atlantic salmon populations in Scotland however a life cycle within the salmon itself had not been characterised (Freeman 2002). Future studies identified a life cycle of *D. lepeophtherii* within the Atlantic salmon and linked *D. lepeophtherii* infections to disease in the salmonid host, which perhaps led to the termination of the patent (Harper 2002; Nylund et al. 2010) (Section 1.8.1.1.). Regardless, this scenario highlights the possibility of using members of this family as biological control agents.

1.10.4 Models for studying eukaryotic extreme genome minimalism

The genome size of *E. bieneusi*, the only species within the Enterocytozoonidae for which there is draft genome available stands at ~6 Mbp (Akiyoshi et al. 2009). Although it is regarded as a small genome, it is almost three times bigger than that recorded for the smallest eukaryotic genome, *Encephalitozoon intestinalis*, 2.3 Mbp (Corradi et al. 2010). The genome of *E. bieneusi*, however, displays more gene loss as compared to that of *Enc. intestinalis* (Keeling & Corradi 2011). That is, whereas the small genome of *Enc. intestinalis* is predominantly as a consequence of deletion of subtelomeric regions, *E. bieneusi*'s moderately small genome is partly due to the absence of most genes responsible for core metabolic processes such as glycolysis, pentose phosphate and trehalose metabolism and fatty acid biosynthesis (Keeling & Corradi 2011). Considering the target host range that spans from vertebrates to invertebrates and from fresh water to marine habitats (Table 1.1), members of the Enterocytozoonidae would be exquisite models for studying eukaryotic genome/metabolic minimalism if this metabolic "incapacity" observed in *E. bieneusi* is indeed a common feature across the Enterocytozoonidae family. In chapter 2 and 3 of this thesis, newly available in-house genomic data generated from environmental samples of members of the Enterocytozoonidae family was mined and whole genome comparative analysis were performed with publicly available genomic data of other members of the microsporidian phylum to investigate if the metabolic "incapacity" displayed by *E. bieneusi* is a common feature within the Enterocytozoonidae family.

1.11 Overall aims and objectives of study

1.11.1 Understand the extent of metabolic loss in the Enterocytozoonidae

Glycolysis is an ATP generating metabolic process which is ubiquitous in most eukaryotic and prokaryotic cells (Fothergill-Gilmore & Michels 1993; Berg et al. 2006). However *E. bienewisi*, a human infecting microsporidian has been recently found to lack this pathway (Akiyoshi et al. 2009; Keeling et al. 2010). Loss of this metabolic process in conjunction with the general absence of oxidative phosphorylation capabilities in the Microsporidia suggests that *E. bienewisi* solely relies on ATP import for its energy needs clusters (Goldberg et al. 2008; Williams et al. 2008). The placement of *Ent. canceri*, a crab infecting intranuclear microsporidia as the closest relative to *E. bienewisi* in phylogenetic studies sparked questions of whether this crab parasite has also lost glycolytic capabilities and if so, how it attained energy from within its unusual intranuclear habitat (Stentiford et al. 2007). In chapter 2, genomic data generated during the course of this Ph. D. will be harnessed to initially ascertain the evolutionary relationship between *E. bienewisi*, *Ent. canceri* and other members of the Enterocytozoonidae sequenced in this study. Following this, in chapter 3, genomic data presented in chapter 2 will be mined to establish if the absence of glycolysis in *E. bienewisi* is a unique feature of this species or a common trait in *Ent. canceri* and other members of the Enterocytozoonidae family. If indeed glycolysis is found to be absent in these genomes, further comparative genomic analyses will be used establish how and when this important metabolic pathway was lost.

1.11.2 Identify signatures of intranuclear living

In chapter 2, it is hypothesised that the intranuclear lifestyle of *Ent. canceri* will reflect in its transporter and effector protein repertoire. Here the genomes of *Ent. canceri*, *E. hepatopenaei*, *Hepatospora* spp. sequenced and assembled during this Ph. D. and those of 23 publicly available microsporidian genomes are mined and compared with the aim of identifying distinctive pathways and or genes that are lost or gained in the genome of the intranuclear parasite. This would be the first assessment of the genome of a eukaryotic parasite that exclusively inhabits an intranuclear niche.

1.11.3 Phylogenetic assessment of the *Hepatospora* genus

Since the discovery of *Hepatospora eriocheir* in Chinese mitten crabs in 2011, two other microsporidia have been found in the pea crab and the edible crab that had SSU regions ~99% identical to that of *H. eriocheir* (Stentiford et al. 2011; Longshaw et al. 2012). This close relationship was rather peculiar as there were distinct differences in karyotype and spore morphology between all three parasites (Stentiford et al. 2011; Longshaw et al. 2012)[Bateman, pers. comm.]. *Hepatospora* spp. are pathogens of commercially important crustaceans and are phylogenetically positioned as the most basal group within the Enterocytozoonidae family (Stentiford, Bateman, et al. 2013). This family is characterised by the presence of prevalent human infecting microsporidian species, *Enterocytozoon bieneusi* and intranuclear infecting species such as *Nucleospora* spp. and *Enterospora* spp. This basal positioning of *Hepatospora* makes this genus an important resource for comparative genomic analyses and in understanding the polarising evolution of genomic and cellular characteristics within this family. Furthermore *Hepatospora* infections have been identified to cause severe mortality in farmed mitten crabs in China thereby posing a serious threat to food security in the region. Considering the importance of this group of parasites, it is critical that the phylogenetic relationships between the above mentioned *Hepatospora/Hepatospora*-looking microsporidia is established so they could be assigned their proper taxonomic names. Considering the current data that shows that microsporidian SSU regions are not ideal for distinguishing between closely related species in phylogenetic analyses, a multi-protein phylogenetic approach is employed in Chapter 4 to assess the relationship between the above-mentioned *Hepatospora/Hepatospora*-looking microsporidia.

Chapter 2 Comparative genomics of the Enterocytozoonidae reveals extreme loss in metabolic capacity

2.1 Introduction

2.1.1 Absence of oxidation phosphorylative pathways in the Microsporidia

The microsporidia are now primary model systems for understanding the process of metabolic, cellular and genomic reduction in eukaryotes with genomes as small as 2.3 Mb encoding as few as 1990 genes (Corradi et al. 2010). A large set of core eukaryotic genes were jettisoned early in the evolutionary history of microsporidia (Nakjang et al. 2013) that left extant microsporidia without a mitochondrial genome and without the ability to generate ATP via the mechanism of oxidative phosphorylation, whilst its fungal relative *Mitosporidium daphniae* retains a mitochondrial genome (Williams et al. 2002; Williams et al. 2008; Haag et al. 2014).

2.1.2 Absence of core metabolic genes in the genome of *Enterocytozoon bieneusi*

Enterocytozoon bieneusi, an important human infecting parasite has however taken minimalism a step further by losing genes responsible for core metabolic pathways such as glycolysis, fatty acid metabolism and the pentose phosphate pathway (Akiyoshi et al. 2009; Keeling et al. 2010). These losses are indeed interesting considering that the Microsporidia are devoid of oxidative phosphorylation pathways (Williams et al. 2008) and were considered to rely on intrinsic glycolysis and ATP import from their host for their energy requirements (Nakjang et al. 2013; Hacker et al. 2014). This makes *E. bieneusi* an exquisite model organism to investigate extreme reduction and parasite-host dependence. *E. bieneusi* belongs to the Enterocytozoonidae family which surprisingly consists of fish and crustacean-infecting parasites and no other human-infecting parasites (Section 1.8) (Stentiford, Feist, et al. 2013). Despite the importance of members of this family as disease agents in humans and economically important aquaculture and fisheries species and their palatability as a model group for reductive genome evolutionary studies, there is currently no known system for propagating these parasites artificially (Desportes et al. 1985; Chilmonczyk et al. 1991; Hedrick et al. 1991; Lom & Dykoá 2002; Stentiford et al. 2007; Stentiford & Bateman 2007; Tourtip et al. 2009; Freeman & Sommerville 2009; Nylund et

al. 2010; Stentiford et al. 2011; Freeman et al. 2013; Palenzuela et al. 2014). Their refractory nature towards artificial propagation and the absence of microsporidian-specific molecular manipulation techniques such as genetic transformation has made direct molecular studies on microsporidian metabolism impossible. With the advent of cheap DNA sequencing, studies investigating metabolism and reductive evolution in this phylum have been based on the phylogenomics and whole genome comparative analysis (Capella-gutiérrez et al. 2012; Cuomo et al. 2012; James et al. 2013; Haag et al. 2014). Even here, this kind of studies have been limited by the fact that *E. bienersi* is the only member of the Enterocytozoonidae family whose genome is publicly available (Akiyoshi et al. 2009; Keeling et al. 2010). Due to its enteric site of infection, *E. bienersi*'s assembled genome was highly contaminated with bacterial sequences consequently making this assembly less favourable for comparative genomic analysis (Nakjang et al. 2013).

In this chapter, I hypothesise that the loss of metabolic capacity observed in the genome of *E. bienersi* is a common trait within its closely related species. To test this hypothesis, genomes of four members of the Enterocytozoonidae (*Enterospora canceri*, *Enterocytozoon hepatopenaei*, *Hepatospora eriocheir* and *Hepatospora eriocheir canceri*) will be sequenced, assembled and compared to that of *E. bienersi* and 18 other publicly available microsporidian genomes. Genomic data for *Hepatospora* spp. presented here was initially used for a separate phylogenomic study presented in chapter 5.

2.1.3 Host nucleus associations in microsporidian infections

In addition to an intimate interaction with the host mitochondria (Section 1.7), some microsporidians also form similar associations with the host nucleus. In *in vitro* *Anncaliia algerae* infections of zebrafish cells, parasite spores form a dense cluster around host nuclei (Monaghan et al. 2011). Also, in infections of human pathogenic microsporidia like *T. hominis* and *E. bienersi*, a meront-host nucleus association is observed (Hollister et al. 1996; Vávra & Larson 1999). A noteworthy observation made by Vávra and Larson (Vávra & Larson 1999), is the nuclear invagination created by the tight association of *E. bienersi* with the host nucleus. The authors also mention that there have been occasions where the developing meront was observed to have partially enveloped itself within the host nucleus however, it has never been observed entirely within the host nucleus. This host

nucleus-parasite association is a particularly uncharted research area that may hold key answers to questions regarding microsporidian metabolism and evolution.

The case of *E. bieneusi*'s association with host nuclei becomes even more compelling when one considers that some of the microsporidian species that are phylogenetically closely related to *E. bieneusi* are intranuclear parasites (Stentiford & Bateman 2007). Examples of these intranuclear lineages include the *Nucleospora* genus whose merogonial stage develops within the nuclei of their piscine hosts (Freeman et al. 2013). Also, the recently described *Enterospora canceri* has been reported to develop entirely within the nucleoplasm of their decapod hosts (Stentiford et al. 2007). The case of *Ent. canceri* is rather peculiar as it is the closest relative of *E. bieneusi* on rDNA based phylogenetic trees (Stentiford et al. 2007). This may mean that there is a similar loss glycolytic capabilities in this species as observed in *E. bieneusi*. If this is true, it raises the question of how *Ent. canceri* obtains energy from within its host nuclei; an environment physically walled off from host mitochondria, the main source of ATP within the host cell. To answer this question, the genomic data of the intranuclear parasite presented in this study, *E. canceri* will be mined and compared to cytoplasmic infecting species to begin to unravel a genomic signature for intranuclear living. I hypothesise that due to its intranuclear lifestyle, the genome of *Ent. canceri* will encode a transporter and effector protein repertoire that is distinct from that of cytoplasm-infecting microsporidia.

2.1.4 Assembling Next Generation Sequencing (NGS) data

Whole Genome Shotgun projects (WGS) within the past decade have evolved from an era of single, long-read, low-throughput Sanger sequencing to long, paired-end read, high-throughput Next Generation Sequencing (NGS) (Sanger & Coulson 1975; Bentley et al. 2008; Imelfort 2009). More recent developments in the field have also seen the emergence of single-cell genome sequencing (Rhoads & Au 2015). This change has resulted in the drastic increase of sequencing speed and reduction in sequencing cost and an increase in the diversity of organisms that can be sequenced, but has also presented significant challenges (Mardis 2006; Mardis 2008; Shendure & Ji 2008; Mardis 2009). One of these problems include the development of assembly programs that can

handle the high data volumes produced by NGS and can also concatenate NGS sequences into a contiguous sequence that reflects the organism's original DNA. The challenge here lay in the fact that traditional overlap-layout-consensus (OLC) assembly approaches which were popular for assembling Sanger sequences at the time, simply joined reads by finding overlaps between them (Myers 1995). This approach was not particularly useful for assembling NGS data because overlap graph building employed by OLC approaches is a slow process and would have therefore been impractical for the assembly of the millions/billions of reads presented by NGS (Compeau et al. 2011). Moreover OLC approaches were unable to handle the relatively high error rates presented by NGS (Myers 1995; Nagarajan & Pop 2013). OLC approaches also functioned on the principle that a pair of reads belonged to the same genomic region if they had an identical overlapping sequence. Even though this principle is untrue as repetitive sequences are common in eukaryotic genomes, it did not matter since Sanger sequences are characteristically long and therefore will often be longer than the repetitive region itself thereby enabling its correct placement in the assembly. NGS sequences are however short and often do not span the entire repetitive genomic region thereby making their assembly with OLC approaches problematic (Nagarajan & Pop 2009).

De Bruijn graphs were among the most widely accepted novel approaches developed to tackle these problems (Chaisson et al. 2004). Here, instead of joining reads by finding best overlapping alignments between them, reads are further fragmented into shorter k-mers and these k-mers are concatenated if they have an overlap of k-1. As such after each correct overlap the contiguous sequence is extended by 1 nucleotide. The Eulerian cycle employed here meant that this could be performed at a relatively faster speed as opposed to OLC approaches that use Hamiltonian cycles (Compeau et al. 2011). Recent assembly programs based on de Bruijn graph building have been optimized to harness the gap-information between paired-end reads to assemble reads in the right orientation and to assemble highly repetitive genomic regions. Four of the most recently widely used de Bruijn graph-based assembly programs were used in this study to perform a *de novo* assembly of the genomic DNA of three members of the microsporidian Enterocytozoonidae family. Although these assemblers, VELVET (Zerbino & Birney 2008), RAY (Boisvert et al. 2010),

SPADES (Bankevich et al. 2012) and A5-MISEQ (Coil et al. 2015) were based on de Bruijn graph building, they differed in the pipelines they employed.

2.2 Main aims of study

- Sequence and assemble genomes of *Ent. canceri*, *E. hepatopenaei* and *H. eriocheir canceri*.
- Construct an updated phylogeny of the Microsporidia following a multi-gene approach
- Understand the extent of metabolic loss in the Enterocytozoonidae
- Identify signatures of intranuclear living

2.3 Methods

2.3.1 Sampling of edible crabs

European edible crab adults (*Cancer pagurus*) were purchased from local fishermen in Weymouth, UK (50°34'N, 2°22'W) in January 2013.

2.3.2 Identification of edible crab infected tissues

The heart, gonad, gill, muscle and hepatopancreas of collected edible crabs that had been previously anaesthetized on ice for 30 minutes were collected into separate containers. Tissues were fixed for 24 hours in 10 % Davidson's seawater fixative (Fredenburgh et al. 2008) and subsequently transferred to 70 % industrial methylated spirit. Identification of infected tissues was done by collaborators at the Centre for Environment, Fisheries and Aquaculture Science (CEFAS) by using confocal and Transmission Electron Microscopy (TEM) imaging as described in (Stentiford, Bateman, et al. 2013).

2.3.3 *Enterospora canceri* and *Hepatospora eriocheir canceri* spore isolation from the edible crab

The hepatopancreases isolated from infected crabs were crushed with a sterile pestle and mortar in 1 x phosphate buffered saline (PBS) solution. The homogenous mash was then filtered through a 100 µm mesh followed by cell sieving through 40 µm filter. The filtrate was topped up with 1 x PBS/triton X-100 (0.1 %) to 50 ml and pelleted at 3220 x g at 4 °C for 10 minutes. The supernatant was then completely removed and the pellet was resuspended in 2.5 ml of water on ice. A Percoll density gradient was created in a 50 ml Falcon tube by the addition 10 ml of progressive Percoll percentages on top of each other in the following order 100-75-50-25 %. The homogenate was slowly added to the top of the gradient and centrifuged in a 4 °C precooled centrifuge at 1500 x g for 45 minutes. Spores collected at the interface of the Percoll layers were washed four times in sterile water before storing them at -20 °C.

2.3.4 *Thelohania* sp. spore isolation from European crayfish claws

Frozen claws of the European crayfish (*Austropotamobius pallipes*) provided by Prof. Grant Stentiford (CEFAS, Weymouth) were thawed on ice for 30 minutes. The exoskeleton of the claws was carefully opened to liberate the musculature, which contained numerous bulbous whitish nodes. *Thelohania* sp. spores were filtered from the isolated muscle tissues following steps outlined in Section 2.3.3

2.3.5 Acquiring spores of *Enterocytozoon hepatopenaei*

Three 250 ml aliquots of prefiltered *E. hepatopenaei* spores suspended in 70 % ethanol were provided by Dr. Ornchuma Itsathitphaisarn (Mahidol University, Bangkok). On receipt, spores were passed through a Percoll purification gradient as detailed in Section 2.3.3

2.3.6 Genomic DNA extraction for sequencing

Aliquots of purified spores were mixed with 20 ml liquid nitrogen in a sterile mortar. The mixture was slowly stirred until it solidified and then ground with the aid of a sterile pestle for 10 minutes. The powder was resuspended again in 20 ml liquid nitrogen, stirred until it solidified and ground for 10 minutes. This step was repeated three times before resuspending the resulting powder in 800 µl of phenol (pH 8.0). The homogenate was transferred to an Eppendorf tube and mixed by inversion and subsequently centrifuged for 10 minutes at 10,000 x g. The recovered aqueous layer was mixed with 400 µl of chloroform and centrifuged for another 10 minutes at 10,000 x g. The aqueous supernatant containing genomic DNA was then passed through a standard ethanol precipitation protocol (Ausubel et al. 2002) and delivered to Exeter sequencing service, UK for library preparation and Illumina sequencing.

2.3.7 Assembling microsporidian genomes

2.3.7.1 Preliminary assemblies

Illumina reads received from Exeter Sequencing Service were analyzed using a quality control tool called FASTQC. With the help of a filtering program called PRINSEQ, FASTQC results were used to identify and filter/trim reads with poor quality scores. The filtered/trimmed reads were used for the assemblies of the four microsporidian genomes with the following programs: VELVET (v1.2.10) (Zerbino & Birney 2008), RAY (v2.3.1) (Boisvert et al. 2010), SPADES (v2.5.1) (Bankevich et al. 2012) and A5-MISEQ-LONGREAD (v3) (Tritt et al. 2012) using protocols outlined in Section 2.3.7.2.1-4. The quality of these preliminary assemblies was then assessed with the command line version of the genome assembly quality assessment program QUILT (v2.3) (Gurevich et al. 2013).

2.3.7.2 Assembly optimization by Illumina read GC content filtering

A GC-content filtering approach was therefore employed to remove reads that may have originated from host or bacterial DNA. To optimize the assemblies, a

progressive GC content cut-off [27 %, 30 %, 33 %, 36 %, (42 % in the case of *Ent. canceri*)] was performed on the raw Illumina reads prior to feeding them again to the four assembly programs. A command line version of a Next Generation Sequencing filtering program called PRINSEQ (Schmieder and Edwards, 2011) was used for this purpose. The command line script used is outlined below:

```
$ perl prinseq-lite.pl -fastq
Hepatospora_TCGAAG_L001_R1_001.fastq.filtered.fastq -fastq2
Hepatospora_TCGAAG_L001_R2_001.fastq.filtered.fastq -
out_format 3 -min_gc [GC cut off=27,30,33,36] -max_gc 100
```

Following this, a bash script was used to sort filtered reads into corresponding forward and reverse reads and to bin orphan reads into a separate file:

```
#!/bin/sh
mkfifo tmp
awk 'NR%4==1{n=$1}NR%4==2{s=$1}NR%4==0{print n,s,$1}'
forward_read.fq | sort -S 2G > tmp & awk
'NR%4==1{n=$1}NR%4==2{s=$1}NR%4==0{print n,s,$1}'
reverse_read.fq | sort -S 2G | join -a1 -a2 tmp - | awk
'NF==5{print $1"\n"$2"\n+\n"$3 >"shuffled.read1.fq";print
$1"\n"$4"\n+\n"$5 >"shuffled.read2.fq"}NF==3{print
$1"\n"$2"\n+\n"$3>"shuffled.OrphanReads.fq"}'
```

2.3.7.2.1 SPADES Assembly

A SPADES assembly was performed for each set of GC-filtered reads. The unix script used for this is as follows:

```
$spades.py --careful -k 33,55,77,99,127 -1 shuffled.read1.fq
-2 shuffled.read2.fq -o spades_resultsGC
```

2.3.7.2.2 A5 MISEQ Assembly:

Here, the miseq_longread A5 pipeline was used to assemble each set of filtered reads. The unix script used for this is as follows:

```
$a5_pipeline.pl lib_file ./A5Assembly
```

Where lib_file is a text file in the following format:

```
[LIB]
```

```
p1= shuffled.read1.fq
p2= shuffled.read2.fq
up= shuffled.OrphanReads..fq
ins=450
```

Where, p1 and p2 represent the location of the forward and reverse reads respectively and “up” is the location of the orphan reads and “ins” is the Illumina paired-end read insert size. This was repeated for each of the GC filtered Illumina read data sets.

2.3.7.2.3 VELVET Assembly

Prior to using Velvet the forward and reverse Illumina files were concatenated using the following Unix script:

```
$velvet_1.2.10/contrib/MetaVelvet-
v0.3.1/shuffleSequences_fastq.pl          shuffled.read1.fq
shuffled.read2.fq shuffled.concatenated.fq
```

The assembly was repeated for each GC content cutoff read dataset with the following command:

```
$velvetg GC/ -cov_cutoff auto -exp_cov auto -unused_reads
yes -very_clean yes -scaffolding no -min_contig_lgth 100 -
ins_length 450 -exp_cov 99
```

2.3.7.2.4 RAY Assembly

The RAY command for the assembly was as follows:

```
$mpiexec -n 1 Ray -k99 -p shuffled.read1.fq
shuffled.read2.fq -s shuffled.OrphanReads.fq -o
./GCRayoutput
```

2.3.7.2.5 Selecting the best preliminary assembly

In order to assess the extent to which each of the GC filtered assemblies represented their respective genomes, open reading frames (ORFs) from these assemblies were queried against a set of 381-core microsporidian proteins (Keeling et al. 2010). The EMBOSS program, GETORF (Rice et al. 2000) was used to extract ORFs from the assemblies.

```
$ getorf -snucleotide1 assembly.fasta -outseq -minsize 100
```

The accession numbers of the 381 core microsporidian proteins in Keeling *et al.* (2010) supplementary_table1 were used to retrieve the FASTA protein sequences with their corresponding microsporidian orthologs from the publicly available database, MICROSPORIDIADB (Aurrecoechea *et al.* 2011). The extracted protein sequences were used to create a BLAST database on our local server with the command line MAKEBLASTDB program (v2.2.28) (Camacho *et al.* 2009):

```
$makeblastdb -in 381proteins.fasta -out database -dbtype
prot
```

The extracted ORFs were then queried against the local 381-protein database using command line BLASTP:

```
$blastp -i Assembly_CDSs.fasta -d 381proteins.fasta -o
CDSsVs381proteins.blastoutput -m 8 -e 1e-05
```

The BLAST hit IDs were extracted using command line CUT command:

```
$cut -f 2 < CDSsVs381proteins.blastoutput >
ORFsVs381proteins.blastoutput.IDs
```

The corresponding *Enc. cuniculi* IDs of the BLAST hit IDs were then extracted using a string of UNIX text editing scripts:

```
$ grep -f CDSsVs381proteins.blastoutput.IDs <
381ProteinsWithOrthogs.list.txt >
CDSsVs381proteins.blastoutput.E_cunOrthologs | cut -f2 <
CDSsVs381proteins.blastoutput.E_cunOrthologs >
CDSsVs381proteins.blastoutput.E_cunOrthologs.E_cunIDs
```

Where file “381ProteinsWithOrthogs.list.txt” is a file retrieved from MICROSPORIDIADB containing all 381 proteins and their corresponding microsporidian orthologs in tab delimited tabular format.

Each corresponding *Enc. cuniculi* ID of the BLAST hits was then used to parse the list of proteins in Akiyoshi *et al.* (2009) to create a .csv file. The following BASH scripts were used:

```
$ sed 's/^/s/g;s/$/%Present%/g' <
CDSsVs381proteins.blastoutput.E_cunOrthologs.E_cunIDs >
CDSsVs381proteins.blastoutput.EcunOrthologs.EunIDs1
$ sed -f <
CDSsVs381proteins.blastoutput.EcunOrthologs.EcunIDs1 <
```



```
proteinsInExcelFile.txt >  
proteinsInExcelFile.presentInAssembly.csv
```

A genome assembly quality assessment program called QUAST (v.2.3) was also used to assess the quality of the assemblies. These assembly statistics together with results from the gene enrichment analysis were plotted on a graph and used to select the best preliminary assembly.

```
$ quast.py ./contigs.fa --eukaryote -o ./quast --min-contig  
200
```

2.3.7.3 Reassembling GC-filtered preliminary assemblies

The raw Illumina reads were mapped back onto best preliminary assembly for each species with the Burrows-Wheeler transformation aligner (BWA) following the user manual. The alignment quality control program, QUALIMAP was then used to assess the Illumina-read coverage for each of the assembled contigs. Based on these results, contigs with low coverage were removed and the remaining contigs were used as the final assembled genome.

2.3.7.4 Comparing predicted protein sets in the genomes of *Hepatospora eriocheir* and *Hepatospora eriocheir canceri*

The genome assembly evaluation pipeline, QUAST (v3.2) (Gurevich et al. 2013) was used to assess the completeness of the assembled genomes of the two *Hepatospora* species. This was performed by initially running QUAST on the assembled contigs of *H. eriocheir canceri* and using the *H. eriocheir* assembly as the reference genome. This was later repeated with *H. eriocheir* being the query genome and *H. eriocheir canceri* being the reference genome. The alignment program, NUCMER (Kurtz et al. 2004) used by the QUAST pipeline is set by default to select the longest and best aligning contig in the case where more than one contig from the query genome aligns to the same region of the reference genome. This feature was exploited to identify repetitive regions unique to each genome. This was done by re-running QUAST on the contigs that were predicted not to align to the reference genome on the initial QUAST run. Predicted aligned and unaligned contigs were graphically displayed with Artemis (Carver et al. 2008) to estimate length (bp), GC content and contig numbers.

In order to determine whether the unaligned contigs of both assemblies were contaminants, they were queried against the NCBI nucleotide database using the command line version of BLASTN (Camacho et al. 2009) and the output was fed

to a taxonomy viewing program, KRONA (Ondov et al. 2011). Since MAKER predicted proteins are tagged with their contig ID, they were easily assigned to their corresponding contigs by the use of command line text editing tools. Differences in the number of predicted proteins between the two species was assessed with the synteny viewer ACT and various command line text editing tools (Carver et al. 2008).

The orthology prediction program ORTHOMCL was run on the protein datasets with a 1e-05 BLASTP e-value cutoff. In order to understand the reason behind the different number of ORFs predicted for the two *Hepatospora* species, the non-orthologous ORFs of each species were queried against the genome of the other species using the BLASTN program with an e-value cutoff of 1e-05. Non-orthologous ORFs which did not align to the genome of the other species were retrieved. I considered a scenario where contigs in one species happened to be longer than syntenic contigs in the other species. In this case, genes predicted on the extended length for the species with the longer contig will not retrieve a blast hit in the other species. To retrieve these genes, contig sets of the two genomes were first aligned against each other with the synteny mapping tool, R2CAT (Husemann & Stoye 2010). Genes predicted to be on syntenic contigs were retrieved. Alignments were viewed with the synteny viewing tool, ACT (Carver et al. 2008).

To estimate the number of unpredicted proteins in the *H. eriocheir* genome, the predicted proteins of *H. eriocheir canceri* were queried against the predicted protein dataset of *H. eriocheir* with BLASTP, and an e-value and percentage identity cutoff of 1e-05 and 70%.

2.3.8 Assembled genome of *Hepatospora eriocheir*

Dr. Bryony Williams of Exeter University, UK provided the assembled genome of *Hepatospora eriocheir*.

2.3.9 MAKER genome annotation

2.3.9.1 Masking repetitive regions

Repetitive elements in the newly sequenced genomes were masked at the beginning of the MAKER annotation pipeline in order to prevent erroneous downstream gene predictions due to false homology with repeat sequences of non-homologous genes. This was performed with the REPEATMASKER and

REPEATRUNNER which masks low and high complexity repeats respectively (Cantarel et al. 2008).

2.3.9.2 *Ab-initio* gene calling

Since microsporidian genomes are characterised by an abundance of species-specific proteins of unknown function, the *ab-initio* gene finding option within the MAKER pipeline (Cantarel et al. 2008) was employed to detect such coding regions. Moreover, an *ab-initio* gene finding approach was imperative since microsporidian genomes display an accelerated rate of evolution and high sequence dissimilarity between homologous genes thereby making BLAST-based gene finding problematic. The SNAP *ab-initio* gene prediction program was therefore separately installed and configured to allow it to be incorporated into the MAKER pipeline. Considering the phylogenetic distance between the Enterocytozoonidae and the Encephalitozoonidae, the *Encephalitozoon cuniculi* gene model file provided with the SNAP gene prediction package was unsuitable for *ab-initio* gene prediction training. Instead, SNAP was trained on the *Enterocytozoon bieneusi* protein set downloaded from MICROSPORIDIADB (Aurrecochea et al. 2011). MAKER was subsequently run with the trained SNAP gene prediction program and the resulting gff3 file was used to retrain SNAP to further improve *ab-initio* gene prediction.

2.3.9.3 BLAST-based gene calling

Evidence based gene predictions was performed by querying the entire manually annotated protein database from SWISSPROT against the newly assembled genomes using BLASTX. Due to the repeat masking performed at the start of the pipeline, false positives arising from erroneous alignment of long stretches of repeats at this stage are reduced. MAKER generates both *ab-initio* and evidence based predictions to produce consensus gene model files: gff3 and FASTA files.

2.3.10 Formatting annotated genomes for GENBANK submission.

2.3.10.1 Assigning names to MAKER-predicted genes

Since MAKER stores its output files into a hierarchy of nested subdirectory layers in order to improve efficiency, the FASTA-formatted proteins predicted by MAKER were transferred into a single folder and merged into a single file with the following command line scripts:

```
$cp **/*proteins.fasta merged_directory#copy all protein
fasta files from subdirectories into a directory called
merged_directory.
$cd merged_directory#move into the directory called
merged_directory
$cat *proteins.fasta > merged_proteins.fasta#merge multiple
protein fasta files into a single file called
merged_proteins.fasta
```

Once all the predicted proteins had been merged into a single file, they were queried against a locally available SWISSPROT database using the following BLASTP script:

```
$blastp -db uniprot_sprot.fasta -query merged_proteins.fasta
-out merged_proteins.blastp -evalue 0.000001 -outfmt 6 -
num_alignments 1 -seg yes -soft_masking true -lcase_masking
-max_hsps_per_subject 1 -num_threads 20
```

The resulting BLASTP output file was in a tabular format and contained a single BLAST hit for each query protein. e.g.:

```
NODE_1216_0_226#sp|Q8SRY5|RL1_ENCCU ...
NODE_1216_0_227#sp|Q54CN8|TAF13_DICDI ...
```

Where the first column is the query name and the second column is the blast hit name.

2.3.10.2 Formatting MAKER-predicted proteins for SEQUIN submission

On a Machintosh computer, the grep function of the text editor, TEXTWRANGLER was employed to convert the BLASTP output file into a multiple SED file. e.g.:

```
s/NODE_1216_0_226#/NODE_1216 [protein=RL1] [gene=RL1]/g
s/NODE_1216_0_227#/NODE_1216 [protein=TAF13] [gene=
TAF13]/g
```

To achieve this, the following patterns were inserted into the “find” and “replace” Sections of TEXTWRANGLER:

```
Find:
(NODE_[0-9]*_[0-9]*_[0-9]*#)\t.*\|.*\|(.*)[A-Z]*\t.*
Replace:
s/\1/\1 \[protein=\2\] \[gene=\2]/g
```

The resulting multiple SED file was used to convert the merged protein FASTA file (merged_proteins.fasta) into a SEQUIN-acceptable protein file with the following command line script:

```
$sed -f multiple_sed_file < merged_proteins.fasta > merged_proteins.sequin.fasta
```

Sequences within “merged_proteins.sequin.fasta” without a BLASTP hit were annotated as “hypothetical protein” by inserting the following patterns into the “find” and “replace” sections of TEXTWRANGLER:

```
Find:
(>NODE_[0-9]*)_[0-9]*_[0-9]*#.*
Replace:
\1 \[protein=hypothetical protein\]
```

The FASTA file containing the nucleotide sequence of the assembled genome was also formatted into a SEQUIN-nucleotide file.

This was done by removing all descriptors on the description line of the FASTA file except for the node/scaffold/contig name and number. Also the full taxonomic name of the organism in question was appended to the end of the definition line as shown below.

```
>NODE_1216_length_64892_cov_325.327301
is changed to
>NODE_1216 [organism=Hepatospora eriocheir]
```

2.3.10.3 Generation of a preliminary feature table file with SEQUIN

The formatted nucleotide file “org.fsa” (where org is the abbreviation of the taxonomic name of the organism in question) and protein files were submitted to a preinstalled SEQUIN (v15.10) program that processed them into a GENBANK feature table file “org.2.tbl”. Submitter details inserted into SEQUIN was used to generate a template file, which was saved as “org.sbt”. Since the description line of each SEQUIN-formatted protein begun with the protein’s corresponding contig ID, SEQUIN was able to map all proteins onto their respective contigs and use their nucleotide co-ordinates to create a preliminary table file such as the one displayed below.

```
>Feature 1c1|NODE_1216
105 2558 mRNA
product hypothetical protein
```

```

105 2558 CDS
product hypothetical protein
transl_table1
protein_id|NODE_1216_29
2694 5450 gene
gene SYLC
2694 5450 mRNA
product SYLC
2694 5450 CDS
product SYLC
transl_table1
protein_id|NODE_1216_1
>Feature |NODE_1216_1
1918 Protein
product SYLC

```

However, this preliminary table file produced by SEQUIN contained a number of errors. Firstly, SEQUIN erroneously added protein features to the end of the table for each contig. Secondly, SEQUIN did not assign a locus_tag to the annotated proteins. Also, hypothetical proteins were not assigned a gene feature. A series of manual edits were performed using both command line and graphical interface tools to address these errors:

```

$awk '/protein_id\t|NODE_*/ { $0=$0 ", " ++i }1' org.tbl >
org.1.tbl #number protein_id's in descending order. These
numbers will later be used for the assignment of locus tags.

```

File org.1.tbl was further edited using the find and replace GEDIT function on TEXWRANGLER in the following manner:

```

Find: >Feature |NODE_[0-9]*_[0-9]*
[0-9]*[0-9]*\tProtein
product.*\n
Replace: "nothing"

```

```

Find: ([0-9]*[0-9]*)\tmRNA
product\thypothetical protein(
[0-9]*[0-9]*CDS
\t\t\tproduct\thypothetical protein

```

```

\t\t\ttransl_table1
\t\t\tprotein_id)\t.*,[0-9]*)
Replace:
\1gene\n\t\t\tlocus_tag\tLocusTagName_\3\n\1mRNA\n\t\t\t\pr
oduct\thypothetical
protein\n\t\t\tprotein_id\tgnl|BwGroupExeterUni|
LocusTagName_\3\n\t\t\t\ttranscript_id\tgnl|BwGroupExeterUni|
mrna.LocusTagName_\3\2\tgnl|BwGroupExeterUni|
LocusTagName_\3\n\t\t\t\ttranscript_id\tgnl|BwGroupExeterUni|
mrna.LocusTagName_\3

```

```

Find:
([0-9]*[0-9]*\tgene
\t\t\tgene.*)"(
[0-9]*[0-9]*\tmRNA
\t\t\tproduct.*)"(
[0-9]*[0-9]*\tCDS
\t\t\tproduct.*
\t\t\t\ttransl_table1
\t\t\t\tprotein_id).*,[0-9]*)
Replace:
\1\n\t\t\tlocus_tag\tLocusTagName_\4\2\n\t\t\t\tprotein_id\tg
nl|BwGroupExeterUni|
LocusTagName_\4\n\t\t\t\ttranscript_id\tgnl|BwGroupExeterUni|
mrna.LocusTagName_\4\3
gnl|BwGroupExeterUni|
LocusTagName_\4\n\t\t\t\ttranscript_id\tgnl|BwGroupExeterUni|
mrna.LocusTagName_\4

```

```

Find: ([<0-9]*[>0-9]*)\tmRNA
Product\thypothetical protein(
[<0-9]*[0-9>]*\tCDS
\t\t\tproduct\thypothetical protein
\t\t\t\ttransl_table1
\t\t\t\tprotein_id).*,[0-9]*)

```

Replace:

```
\1gene\n\t\t\tlocus_tag\tLocusTagName_\3\n\1mRNA\n\t\t\t\product\thypothetical  
protein\n\t\t\tprotein_id\tgnl|BwGroupExeterUni|  
LocusTagName_\3\n\t\t\ttranscript_id\tgnl|BwGroupExeterUni|  
mrna.LocusTagName_\3\2\tgnl|BwGroupExeterUni|  
LocusTagName_\3\n\t\t\ttranscript_id\tgnl|BwGroupExeterUni|  
mrna.LocusTagName_\3
```

Find:

```
([<0-9]*[0-9>]*\tgene  
\t\t\tgene.*)(  
[<0-9]*[0-9>]*\tmRNA  
\t\t\tproduct.*)(  
[<0-9]*[0-9>]*\tCDS  
\t\t\tproduct.*  
\t\t\ttransl_table1  
\t\t\tprotein_id).*,([0-9]*)
```

Replace:

```
\1\n\t\t\tlocus_tag\tLocusTagName_\4\2\n\t\t\tprotein_id\tgnl|BwGroupExeterUni|  
LocusTagName_\4\n\t\t\ttranscript_id\tgnl|BwGroupExeterUni|  
mrna.LocusTagName_\4\3\tgnl|BwGroupExeterUni|  
LocusTagName_\4\n\t\t\ttranscript_id\tgnl|BwGroupExeterUni|  
mrna.LocusTagName_\4
```

2.3.10.4 Generation of final sequin file for GENBANK submission

The following files were remotely transferred from the Machintosh platform into a single folder in a UNIX computer that had the GENBANK command line tool TBL2ASN pre-installed.

1. org.tbl #final feature table file
2. org.fsa #assembled contigs in FASTA format
3. org.sbt #template file created by SEQUIN
4. org.txt #file containing assembly and coverage data

This was because the Machintosh version of TBL2ASN available on the GENBANK website had a bug and was unable to run. The TBL2ASN program was used to create the final sequin file with its corresponding error files with the following script:

```
$tbl2asn -p ./ -t org.sbt -M n -Z discrep -V -V b
```

The -V b option was to create a GENBANK formatted annotation file which would later be used to identify ORFs missed by MAKER.

2.3.10.5 Identification of ORFs missed by MAKER

A combination of the genome viewer, ARTEMIS (Carver et al. 2008), an ORF calling tool, GETORF (Rice et al. 2000) and BLASTP were used to identify and annotate ORFs that were missed by MAKER. New annotations were converted into a GENBANK feature table by following protocols in Section 2.3.10.3. The resulting feature table was copied and pasted into org.tbl (The feature table created with MAKER annotation). TBL2ASN was rerun in the new feature table to create the final SEQUIN file submitted to GENBANK.

Prior to uploading the resulting SEQUIN file onto GENBANK, the accompanying error files were parsed and all flagged residual formatting errors were addressed.

2.3.11 Identification of gene families

The 21 orthologous proteins used in this analysis (Wrs1p, Taf10p, TFIIE, Taf10p, Tfa2p, Sec62p, Abd1p, SPT16, Brn1p, Nob1p, Caf40p, Tfb2p, Bos1p, Npl4p, Tma46p, Tfa1p, Clp1p, Spt5p, Sec63p, Pri2p and Enp2p) were selected as they had been previously used to generate a multi-protein phylogeny by Nakjang *et al.* (2013). Microsporidian homologous proteins were identified by doing a All-vs-All local BLASTP search (Mount 2007). This BLAST search involved the pooling of predicted proteins from the newly sequenced genomes and those retrieved from MICROSPORIDIADB public webserver (Aurrecoechea et al. 2011) into a single FASTA file. This file was subsequently formatted into a BLASTable database with the MAKEBLASTDB command line tool (Camacho et al. 2009) as explained in Section 2.3.7.2.5. Finally the pooled list of proteins was queried with BLASTP (Camacho et al. 2009) against their own database with an e-value cutoff of 1e-03. The BLASTP output was passed to the command line tool, ORTHOMCL (Li et al. 2003), which clustered the proteins into homologous groups.

2.3.12 Phylogenomic assessment of the microsporidian phylum

2.3.12.1 Concatenated protein alignment

Each of the 21 protein sets was aligned with command line version of MUSCLE (v3.8.31). The most rigorous masking option on TRIMAL (v1.2rev59) (Capella-Gutierrez et al. 2009) was employed to mask each alignment by using the “-strictplus” option. The best substitution models for each of the 21-masked protein alignment was predicted with the PROTTEST (v3.4.2) command line tool (Abascal et al. 2005) by following the user guide. These substitution models were passed to the command line version of RAXML (v8.2.8) (Stamatakis 2014) with the “-m” option to perform maximum likelihood phylogenetic analysis on each of the masked protein alignment. This process was important to identify and remove unlikely orthologs, which usually form long branches in phylogenetic trees. Following the replacement of spurious orthologous proteins and the absence of unusual branches was ascertained; the individual masked protein alignments were manually concatenated using SEAVIEW (v4) (Gouy et al. 2010).

2.3.12.2 Maximum likelihood analysis on 21-protein concatenated alignment

Finally, the concatenated phylogenetic tree was constructed with RAXML using a GTR+GAMMA substitution+rate heterogeneity model (Stamatakis 2014). In this final multi-protein concatenated tree, the variable rates of substitution for the individual protein sets were taken into account by using the “-q” parameter to call on a partition file containing the coordinates of each protein set and their predicted substitution model. See Appendix 4 for details of the partition file and the command line script employed for this analysis.

2.3.12.3 Bayesian inference analysis on 21-protein concatenated alignment.

To corroborate the phylogenetic relationships estimated by maximum likelihood analysis, the concatenated alignment was also used to construct a consensus tree with the Bayesian inference method. Here, a command file containing the coordinates of each protein set (partitions) in the alignment was manually created to enable the MRBAYES program to analyse each partition with a different evolutionary model (See Appendix 4). The Bayesian inference analysis executed 20000 generations, which were sampled after every 300th time. By default, a

Metropolis-coupled Markov Chain Monte Carlo Method was used to run 4 chains in parallel to permit for a more thorough exploration of the data. Although 20000 generations were initially executed, 36000 generations were needed in this analysis to achieve a standard deviation of split frequencies between the two independent runs below 0.01. The potential scale reduction factor (PSRF) values in the summary table were close to 1 for all parameters. The burn-in fraction of 25 % was used to obtain the consensus phylogram and posterior probabilities for each bipartition using the sumt command.

2.3.13 Comparative genomic analysis

2.3.13.1 Genomes used in this study

Comparative genomic analysis included new genomes sequenced in this study: *Enterospora canceri*, *Enterocytozoon hepatopenaei*, *Hepatospora eriocheir* (Sequenced genome provided by Dr. Bryony Williams), *Hepatospora eriocheir canceri* and *Thelohania* sp., which are all, kept on an in-house server and have also been submitted to the GENBANK online repository (*Thelohania* sp. has not been submitted to GENBANK on collaborator's request). The protein sets for remaining microsporidians used in this study: *Anncaliia algerae*, *Ordospora colligata*, *Trachipleistophora hominis*, *Spraguea lophii*, *Vittaforma corneae*, *Encephalitozoon romaleae*, *Vavraia culicis*, *Edhazardia aedis*, *Encephalitozoon hellem* Swiss, *Encephalitozoon hellem* ATCC, *Nematocida parisii* ERTm1, *Nematocida parisii* ERTm3, *Nematocida* sp. ERTm2, *Nematocida* sp. ERTm6, *Enterocytozoon bieneusi*, *Encephalitozoon intestinalis*, *Encephalitozoon cuniculi* and *Nosema ceranae* were retrieved from publicly available database, MICROSPORIDIADB (Aurrecoechea et al. 2011). Protein sets for *Rozella allomycis* and *Mitosporidium daphniae* were downloaded from online publicly available NCBI records (Tatusova et al. 2014).

2.3.13.2 Identifying core microsporidian proteins in the newly sequenced genomes

In order to assess how the protein repertoire coded by the genomes sequenced in this study compared to that of other microsporidians, the predicted open reading frames (ORFs) of all four Enterocytozoonidae genomes were compared against a set of 381-core microsporidian proteins used for a similar study in Keeling et al., 2010. To achieve this, proteins coded by the 19 publicly available

microsporidian genomes were initially retrieved as FATSA files from the MICROSPORIDIADB database (Aurrecochea et al. 2011). Proteins from all microsporidians were then pooled into a single FASTA file and converted into a BLAST database using command line MAKEBLASTDB tool, included in the BLAST+ package (Camacho et al. 2009). BLASTP was used to query the pooled proteins against their own database (All-vs-All BLAST search). Orthologous protein clusters were subsequently identified by parsing the tabular formatted BLASTP output with a Markov cluster algorithm known as ORTHOMCL (Li et al. 2003) set on the following parameters: MCL inflation=1.1 and maximum weight=100. Finally, orthologous clusters of the 381-core microsporidian proteins were extracted from the ORTHOMCL output and manually formatted into an Excel file.

2.3.13.3 Identifying putative transcription factor binding domains

The online FEATUREEXTRACT server (v1.2) (Wernersson 2005) was used to extract 100 bp nucleotides upstream of ORFs from each of the assembled genomes. Only ORFs with more than 8 nucleotides upstream of their start codons were used. These upstream regions, were passed to the MEME-CHIP online server (Machanick & Bailey 2011), which was set on default parameters. The Eukaryotic DNA parameter was however changed to JASPAR CORE and UNIPROBE Mouse. Only the logos of motifs predictions with e-value higher than 1e-03 and nucleotide length higher than 5 bp were retained.

2.3.13.4 Scanning genomes for transposable elements and tRNAs

The command line version of the DFAMSCAN tool (Hubley et al. 2015) with default parameters was used to scan each of the assembled genomes for likely transposable elements. This was repeated with the REPEATMASKER (v4.0.6) command line tool (Tarailo-Graovac & Chen 2009) configured on RMBLAST (v2.2.28) (Camacho et al. 2009) using a locally installed REPBASE database (Bao et al. 2015). The command line version of TRNASCAN-SE (v1.3.1) (Lowe & Eddy 1997) was used to identify tRNAs in the sequenced genomes by following the user manual.

2.3.13.5 Assessment of synonymous codon usage and codon usage bias within sequenced *Enterocytozoonidae* genomes

The CODONW (Néron et al. 2009) program from the online MOBYLE suite (<http://mobyale.pasteur.fr/cgi-bin/portal.py>) was used to generate the data for the codon usage plots by running it on predicted ORFs from the assembled genomes. These predicted ORFs were also submitted to the online tool CODONO (Angellotti et al. 2007). This program used synonymous codon usage order (SCUO) measurement to estimate the level of codon usage bias for each gene in the sequenced genomes. A Wilcoxon Two Sample Test was used to assess the difference in SCUO distributions of genes between genomes.

2.3.13.6 Identification of transporter proteins

Predicted ORFs from the new genome assemblies of *Ent. canceri*, *E. hepatopenaei*, *H. eriocheir canceri*, *H. eriocheir* and *Thelohania* sp., and the genomes of the 18 microsporidian species mentioned in Section 2.3.13.1 were merged into a single FASTA file and submitted to the following automated processes. The command line version of the transmembrane domain prediction tool, TMHMM (Krogh et al. 2001) was used to select proteins with one or more transmembrane domains. Since the TMHMM program is known to erroneously identify signal peptides as transmembrane domains the protein set was also analysed with SIGNALP (Petersen et al. 2011). Proteins predicted to have a single transmembrane domain that was also identified to be a signal peptide were removed from the analysis. A BLASTP search with the remaining protein set was carried out against the following databases: SGD, TCDB and NCBI databases (Saier et al. 2006; Cherry et al. 2012; Tatusova et al. 2014). Where possible, a consensus BLAST hit ID was selected for each putative transmembrane protein. WOLFPSORT (Horton et al. 2007) was subsequently used to predict a subcellular localization for these proteins. Predicted plasma membrane proteins that had four or more transmembrane domains were selected for further analysis as potential transport proteins.

For each predicted plasma membrane transporter, its orthologous cluster (from Section 2.3.13.2) was manually parsed to recover orthologs that may have been missed by the automated pipeline described above. Although 18 non-*Enterocytozoonidae* microsporidian species were used in this analysis, only data

for *Enc. cuniculi* and *T. hominis* were displayed in the final results. Appendix 5 contains more details on the bash scripts employed in this section.

2.3.13.7 Identification of secreted proteins

Predicted ORFs from the newly sequenced genomes and those from published genomes were parsed with the TMHMM (Krogh et al. 2001) program to detect proteins with transmembrane domains. Proteins predicted to be devoid of transmembrane domains or only possess a single transmembrane domain were retained for further analysis. Proteins predicted to have a single transmembrane domain were retained because the TMHMM program is known to erroneously predict signal peptides as transmembrane domains. This protein set was passed to the signal peptide prediction program, SIGNALP (Petersen et al. 2011). This program scans the first 70 amino acids of proteins to detect signal peptides. Proteins predicted by this program to be devoid of transmembrane domains but possess a signal peptide were retained as the putative secretome of the microsporidian species being analysed. The ORTHOMCL orthologous protein cluster file from Section 2.3.13.2 was used to assess orthology of the predicted secreted proteins. Finally, a Gene Ontology (GO) assessment of the predicted secreted proteins was performed with the graphical interphase version of BLAST2GO 3.2 (Conesa et al. 2005). Bash scripts employed in this section can be found in Appendix 7.

2.4 Results

2.4.1 Assembly and optimization of *Hepatospora eriocheir canceri* genome

2.4.1.1 Preliminary assembly

A total of 8,740,870 reads with a mean length of 450 bp were produced by the Exeter Sequencing service. The preliminary assembly performed with four eukaryotic genome assemblers (VELVET, RAY, SPADES, A5-MISEQ) produced a total assembly size ranging between 47 to 103 Mbp with the A5-MISEQ assembly presenting the highest N50 value of 706. Statistics for the preliminary assembly performed with these reads are summarized in Table 2.1 below.

Table 2.1: Comparing the performance of four eukaryotic genome assembler on the genome of *Hepatospora eriocheir canceri*

Assembly programs	Velvet	RAY	SPADES	A5 -MISEQ
-------------------	--------	-----	--------	-----------

Contigs (>= 0 bp)	292,784	179,954	692,590	61,652
# Contigs (>= 1000 bp)	1,135	3,892	1,514	6,785
Total length (>= 0 bp)	103,560,848	88,992,682	118,251,282	47,984,045
Total length (>= 1000 bp)	1,762,454	9,893,136	4,035,989	12,986,038
# contigs	292,784	179,948	692,590	61,652
Largest contig	14,923	48,516	48,515	48,516
GC (%)	42.03	42.44	41.63	42.04
N50	353	462	171	706
NG50	870	8,314	2,699	5,555
# N's per 100 kbp	0.14	0.06	0	0.63

2.4.1.2 Optimizing assembly by GC-content Illumina read filtering

2.4.1.2.1 SPADES assembly

The total assembly length for the different GC-content cut-offs ranged between 4.4 Mbp to 3.8 Mbp with GC27 cut-off presenting the highest assembly size (Table 2.2).

Table 2.2: Comparing SPADES assemblies at different Illumina read GC content cut-offs for the genome of *Hepatospora eriocheir canceri*

SPADES Assembly				
	GC27	GC30	GC33	GC36
# Contigs (>= 0 bp)	3,810	1,854	1,742	1,814
# Contigs (>= 1000 bp)	1188	1000	883	801
Total length (>= 0 bp)	4,365,493	3,883,903	4,104,910	4,215,459
Total length (>= 1000 bp)	3,040,075	3,482,117	3,732,192	3,801,797
Largest contig	15,823	27,220	36,019	85,723
GC (%)	21.87	22.39	22.8	23.17
N50	1,962	4,305	5,627	6,714
NG50	1,777	3,161	4,495	5,590
# N's per 100 kbp	0	0	0	0

2.4.1.2.2 A5-MISEQ assembly

The total assembly length for the different GC-content cut-offs ranged between 4.2 Mbp and 8.1 Mbp with GC36 cut-off presenting the highest assembly size (Table 2.3).

Table 2.3: Comparing A5-MISEQ assemblies at different Illumina read GC content cut-offs for the genome of *Hepatospora eriocheir canceri*

A5-MISEQ Assembly				
	GC27	GC30	GC33	GC36
# Contigs (>= 0 bp)	2,995	3,428	4,343	6,597
# Contigs (>= 1000 bp)	1468	1554	1593	1746
Total length (>= 0 bp)	4,184,886	5,280,755	6,358,504	8,130,429
Total length (>= 1000 bp)	3,130,621	4,023,687	4,560,798	5,038,479

# contigs	2,995	3,428	4,343	6,597
Largest contig	14,796	20,219	25,690	25,748
GC (%)	21.39	22.76	24.06	26.01
N50	1,769	2,247	2,300	1,636
NG50	1,554	2,633	3,546	4,238
# N's per 100 kbp	0	0	0	0

2.4.1.2.3 VELVET assembly

The total assembly length for the different GC-content cut-offs ranged between 5.8 Mbp and 19.7 Mbp with GC36 cut-off presenting with the highest assembly size (Table 2.4).

Table 2.4: Comparing VELVET assemblies at different Illumina read GC content cut-offs for the genome of *Hepatospora eriocheir canceri*

VELVET Assembly				
	GC27	GC30	GC33	GC36
# Contigs (>= 0 bp)	12,362	19,291	30,403	50,473
# Contigs (>= 1000 bp)	793	1,124	1,319	1,394
Total length (>= 0 bp)	5,782,560	8,494,908	12,568,872	19,666,629
Total length (>= 1000 bp)	1183683	1730090	2096828	2290345
# Contigs	12362	19291	30403	50473
Largest contig	5718	7053	11090	9391
GC (%)	21.59	23.6	25.9	28.65
N50	508	444	378	358
NG50	613	787	913	976
# N's per 100 kbp	0	0	0	0

2.4.1.2.4 RAY assembly

The total assembly length for the different GC-content cut-offs ranged between 7.6 Mbp and 35.6 Mbp with GC36 cut-off presenting with the highest assembly size (Table 2.5).

Table 2.5: Comparing RAY assemblies at different Illumina read GC content cut-offs for the genome of *Hepatospora eriocheir canceri*

RAY Assembly				
	GC27	GC30	GC33	GC36
# Contigs (>= 0 bp)	33,099	67,194	124,164	202,769
# Contigs (>= 1000 bp)	997	995	914	949
Total length (>= 0 bp)	7,585,488	13,438,331	22,775,538	35,587,421
Total length (>= 1000 bp)	2,131,987	2,806,945	3,069,422	3,375,253
# Contigs	33,099	67,194	124,164	202,769
Largest contig	13867	16073	21428	32644
GC (%)	23.12	25.62	28.13	30.47
N50	208	182	173	170

NG50	795	1607	2277	2920
# N's per 100 kbp	0	0	0	0

2.4.1.2.5 Selecting best assembly program for the genome of *Hepatospora eriocheir canceri*

There was a noticeable difference in the assembly outputs of the four assemblers used in this study. Two assemblers, namely, RAY and VELVET produced final assemblies with over 25,000 contigs and an N50 value below 510 (Figure 2.1). The assemblies, at each GC content cut-off of A5-MISEQ and SPADES were consistently below 5000 contigs and had N50 values above 1500 (Figure 2.1). As opposed to the assembly sizes of RAY and VELVET which were over 5.7 Mbp, those of A5-MISEQ and SPADES were consistently below 5.5 Mbp (Figure 2.1), which is in the range of the estimated genome size of *E. bieneusi* (6 Mbp) (Akiyoshi et al. 2009; Keeling et al. 2010) and the assembly size of *Hepatospora eriocheir* (4.7 Mbp) (Data provided by Bryony Williams). In light of these results, A5-MISEQ and SPADES were considered to be more appropriate for microsporidian genomes belonging to the Enterocytozoonidae. These were therefore the only assemblers used for the other microsporidian genomes. In the end, A5-MISEQ was chosen over the SPADES assembler. This was because the SPADES assembly values were difficult to interpret as an expected increase in assembly size with increasing GC content cut-offs was not observed here (Figure 2.1). Instead, the highest assembly size was recorded for GC27 cut-off whereas the lowest assembly size was recorded for GC30 cut-off (Figure 2.1). As such even though SPADES presented assemblies with the highest N50 values, A5-MISEQ was chosen as the best assembly program.

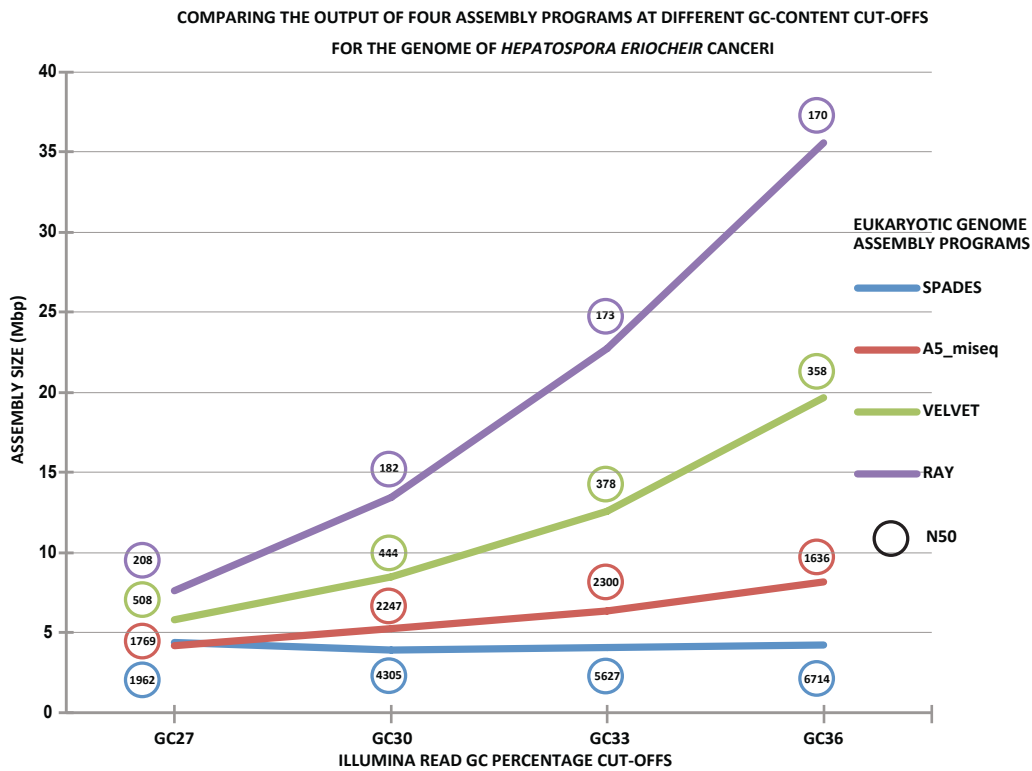


Figure 2.1: Comparing the Assembly sizes and N50 values for the genome of *Hepatospora eriocheir canceri* produced by four eukaryotic genome assemblers at increasing Illumina read GC content cut-offs.

2.4.1.2.6 Selecting the best GC content cut-off for the genome of *Hepatospora eriocheir canceri*

When the completeness of the A5-MISEQ assembled genome at each GC-content cut-off was assessed, there appeared to be a sharp rise in the number of core gene set to 86.8% at GC33 cut-off. The same percentage was recorded for the next GC cut-off and this value only increased slightly to 87.1 % when the raw reads were used without any GC filtering (GC100). The N50 value peaked at GC33 cut-off at 2300 and decreased to 706 with the raw read assembly (Figure 2.2).

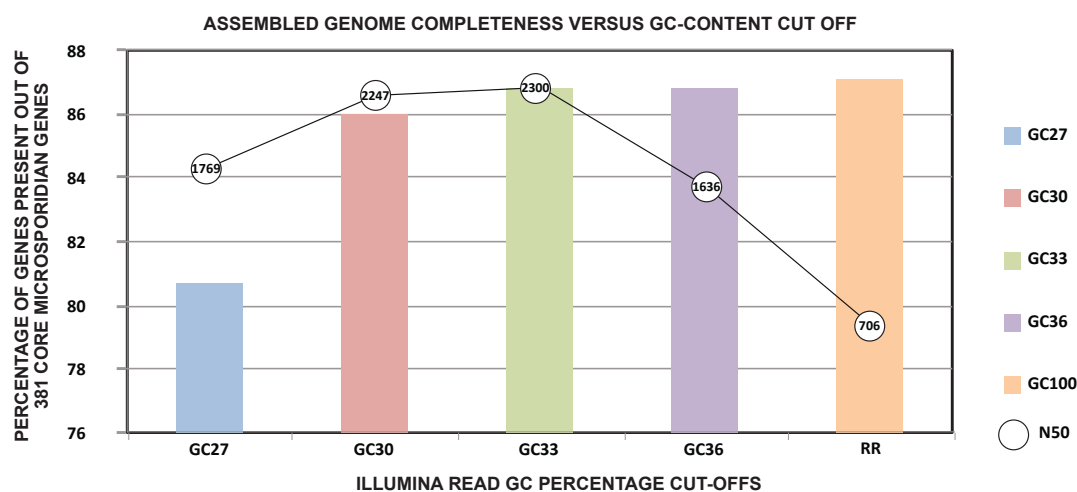


Figure 2.2: Comparing completeness between genome assemblies of *Hepatospora eriocheir canceri* performed with Illumina reads filtered at various GC content cut-offs. RR represents raw read assembly.

2.4.1.2.7 Further assembly optimization by aligning genomes of the two *Hepatospora* spp. to each other

The automated MAKER annotation pipeline predicted a total of 3288 and 2675 proteins from the genomic DNA assembly of *H. eriocheir canceri* and that of *H. eriocheir* respectively. The majority of *H. eriocheir*'s predicted proteins were on contigs that aligned to the genome of *H. eriocheir canceri* (Figure 2.3A). 18 of the remaining proteins from *H. eriocheir* were located on short repetitive contigs whereas 2 were located on contigs that did not align to either repetitive contigs or *H. eriocheir canceri*'s genome. Moreover, these proteins or the contigs they appeared on did not have any hit when queried against the NCBI database (Figure 2.3A). Finally, 20 unaligned *H. eriocheir* contigs spanning 3315 bp had BLASTN hits for microsporidian rDNA when queried against the NCBI nucleotide database (Figure 2.3A).

Whereas the assembled genome length of *H. eriocheir canceri* that aligned to *H. eriocheir* was almost identical, 4.58 and 4.56 Mbp respectively, more proteins were predicted in the aligned contigs of *H. eriocheir canceri* than in its sister subspecies (Figure 2.3A). The majority of the unaligned length (1.52 Mbp) for *H. eriocheir canceri* had no hits when queried against the NCBI nucleotide database. Only 0.02 Mbp of the unaligned length had arthropod BLASTN hits (representing possible host contamination). The repeat length of the *H. eriocheir canceri* genome was 0.26 Mbp and it contained 100 predicted proteins (Figure 2.3A). A Blast2GO run on these 100 proteins annotated 15 as DNA/RNA binding and modification proteins, 1 as a sugar transporter, 1 as nuclear membrane exporter, 1 as ribosomal protein, 1 as compound binding protein and 1 as a protein with

dimerization activity. The remaining 80 proteins had either no BLAST hit or had a BLAST hit to a microsporidian hypothetical protein.

The final assemblies included aligned contigs and unaligned contigs that had a microsporidian hit upon querying against the NCBI database. Statistics of the final assemblies submitted to NCBI are summarized in Table 2.18.

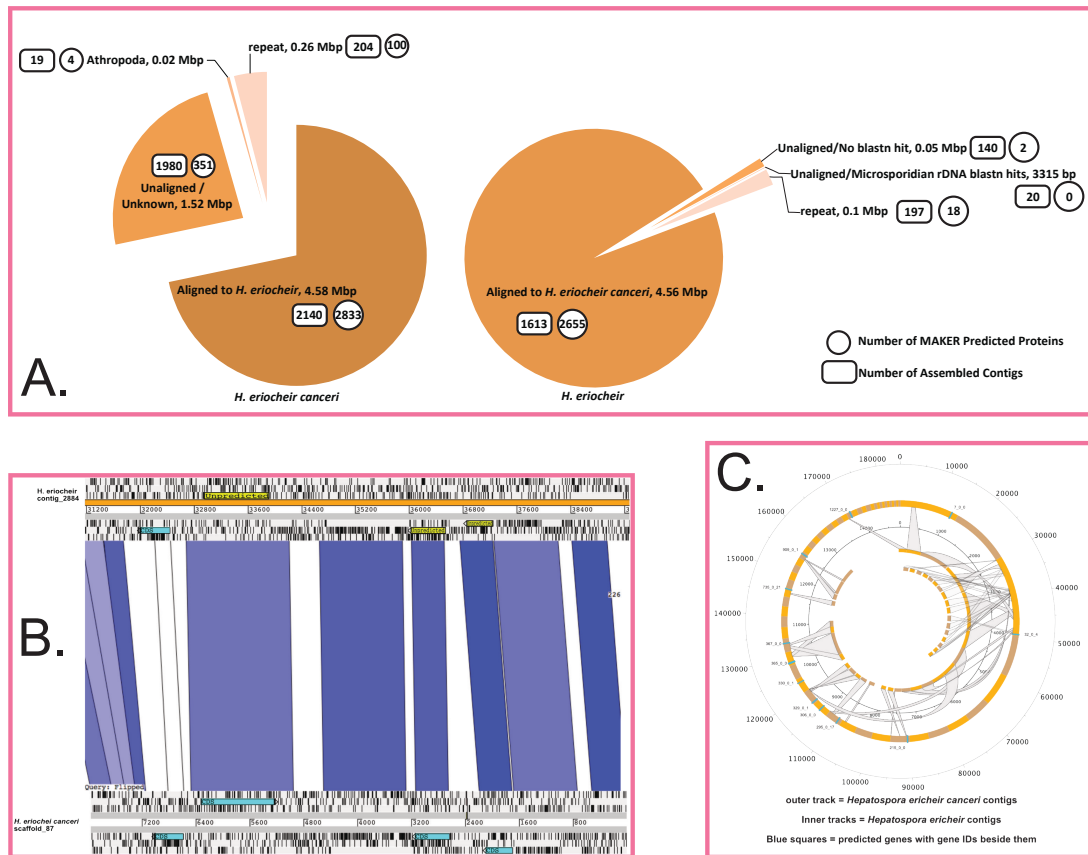


Figure 2.3: Comparing genomic DNA assemblies of the two *Hepatospora* species. A. Shows length of assembly that aligned to the sister species, unaligned length with no BLASTN hits, unaligned length but had BLASTN hits to arthropod sequences and length of repetitive region of the genomes. B. Screen shot from ACT synteny viewer of representative syntenic contigs/scaffolds from the *Hepatospora* species that aligned to each other. This shows that some open reading frames in the genome of *Hepatospora eriocheir* were missed by the MAKER annotation pipeline. C. Representative scaffolds from *H. eriocheir cancri* that mapped onto the genome of *H. eriocheir* but contained genes that had no BLASTN hits when queried against the genome of *H. eriocheir*. This demonstrates that unmapped genes (blue squares) were as a consequence of fragmented contigs in *H. eriocheir* that did not span across gene coding regions.

2.4.2 Assembly and optimization of *Enterocytozoon hepatopenaei* genome

2.4.2.1 Preliminary assembly

A total of 11,332,956 paired-end Illumina reads, with sequence length ranging between 35-300 bp were produced by the Exeter Sequencing service. Subsequent filtering of extremely short reads and trimming of remnant adapter sequences resulted in a total of 7,366,754 paired-end reads ranging between

180-210 bp. These reads were used for filtering steps and assemblies. The preliminary assemblies performed on these Illumina reads with SPADES and A5-MISEQ was 14.25 Mbp and 5.34 Mbp respectively with the A5-MISEQ assembly presenting the highest N50 value of 2870 (Table 2.6).

Table 2.6: Comparing SPADES and A5-MISEQ assemblies performed on the raw Illumina reads of *Enterocytozoon hepatopenaei*

	SPADES	A5-MISEQ
# Contigs (>= 0 bp)	34278	3082
# Contigs (>= 1000 bp)	1552	1144
Total length (>= 0 bp)	14251002	5336795
Total length (>= 1000 bp)	6434187	4068302
Largest contig	61185	118107
GC (%)	36.59	32.05
N50	1908	2870
# N's per 100 kbp	110.94	76.41

2.4.2.2 Optimizing assembly by Illumina read GC-content filtering

2.4.2.2.1 SPADES assembly

Total assembly lengths ranged between 1.91 Mbp to 4.23 Mbp with the highest N50 of 7182 presented at a GC content cut-off of 36 %. The lowest number of assembled contigs, 2017 was however recorded at GC-content cut-off of 33 % (Table 2.7). When predicted open reading frames in each assembly were queried by BLASTP against a set of 381-core microsporidian proteins, a positive correlation was observed between GC-content cut-off and number of core proteins present in each assembly (Figure 2.4A). That is, with increasing GC content filtering, there were a higher number of core microsporidian proteins present in the assembly. However, upon closer inspection of the proteins from each assembly, it was evident that not all BLAST hits were microsporidian proteins. For example, for the assembly performed on Illumina reads filtered at 24 % GC, only 105 out of the 120 BLAST hits were predicted by KRONA (Ondov et al. 2011) as originating from Microsporidia, 3 could not be assigned to any taxa and the remaining 12 belonged to other eukaryotic lineages other than the Microsporidia (Figure 2.4A) (Appendix 1). There was an increase in the number of predicted microsporidian proteins with increasing Illumina read GC percentage filtering until GC 39 % (Figure 2.4A) (Appendix 1). At this point, 4 of the proteins were predicted to have a bacterial origin, 6 of the proteins were predicted to have a eukaryotic origin other than microsporidian and 2 proteins were unassigned to any taxon by KRONA (Figure 2.4A) (Appendix 1).

Table 2.7: Comparing SPADES assemblies at different Illumina read GC content cut-offs for the genome of *Enterocytozoon hepatopenaei*

SPADES Assembly							
GC content cut-off	24	27	30	33	36	39	Unfiltered reads
# Contigs (>= 0 bp)	2243	2095	2061	2017	2366	3011	34278
# Contigs (>= 1000 bp)	509	683	651	549	518	563	1552
Total length (>= 0 bp)	1909527	2409159	2880958	3215995	3656878	4232554	14251002
Total length (>= 1000 bp)	995291	1652823	2186075	2494746	2763749	3022864	6434187
Largest contig	12154	25597	25721	35270	49850	103292	61185
GC (%)	22.05	23.68	24.9	25.98	27.25	28.86	36.59
N50	1059	1811	3289	5353	7182	6111	1908
# N's per 100 kbp	24.31	20.44	25.06	46.19	47.03	31.51	110.94

2.4.2.2.2 A5-MISEQ assembly

Total assembly length ranged between 1.66 Mbp to 3.75 Mbp with the highest N50 of 6928 presented at a GC-content cut-off of 39 %. The lowest number of assembled contigs, 1473 was however recorded at GC-content cut-off of 33 % (Table 2.8). When predicted open reading frames in each assembly were queried by BLASTP against a set of 381-core microsporidian proteins, a positive correlation was observed between GC-content cut-off and number of core proteins present in each assembly (Figure 2.4B). That is, with increasing GC-content filtering, there were a higher number of core microsporidian proteins present in the assembly. However, upon close inspection of the proteins from each assembly, it was evident that not all proteins were microsporidian. For example, the assembly performed on Illumina reads filtered at GC 24 %, only 80 of the 123 ORFs were predicted by KRONA as originating from Microsporidia, 32 could not be assigned to any taxa, 1 was predicted to have bacterial origin and the remaining 10 belonged to other eukaryotic lineages other than the Microsporidia (Figure 2.4B) There was an increase in the number of KRONA-predicted microsporidian proteins with increasing Illumina read GC percentage filtering similar to that observed in the SPADES assembly.

Table 2.8: Comparing A5-MISEQ assemblies performed with Illumina paired-end reads filtered at different GC percentage content of *Enterocytozoon hepatopenaei*

A5-MISEQ							
GC content cut-off	24	27	30	33	36	39	Unfiltered reads
# Contigs (>= 0 bp)	1753	1838	1600	1472	1527	1841	3082
# Contigs (>= 1000 bp)	476	700	716	607	569	591	1144
Total length (>= 0 bp)	1661748	2254717	2690586	3010074	3332823	3753215	5336795
Total length (>= 1000 bp)	863767	1524855	2131339	2463251	2730829	2973923	4068302
Largest contig	9272	25589	25716	27420	49845	103635	118107
GC (%)	21.52	23.25	24.54	25.5	26.47	27.74	32.05
N50	1047	1571	2890	4825	6794	6928	2870
# N's per 100 kbp	85.09	77.3	70.95	74.95	73.48	77.75	76.41

2.4.2.3 Selecting best assembly program for the genome of *Enterocytozoon hepatopenaei*

When the predicted ORFs from each assembly were queried by BLASTP against a set of 381-core microsporidian proteins, it was evident that the non-filtered reads assembled by the A5-MISEQ assembler produced the most complete assembly as it produced the maximum number of BLASTP hits (Figure 2.4B RR). Secondly, this assembly appeared to possess fewer contaminant hits as opposed to its analogous SPADES assembly (Figure 2.4A RR). Since the most complete assembly was attained with the unfiltered Illumina reads, further filtering steps on the contigs assembled with the unfiltered reads were employed to remove contamination.

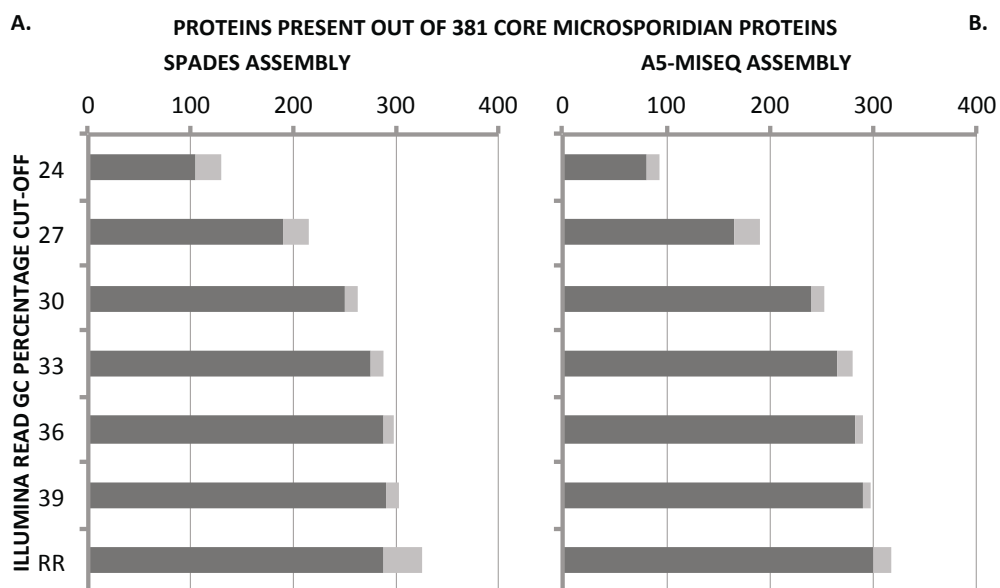


Figure 2.4: Estimating completeness of assembled genome of *Enterocytozoon hepatopenaei* at different GC content cut-offs. Deep grey = KRONA predicted microsporidian proteins. Light grey = proteins predicted as non-microsporidian or not assigned to any taxa by KRONA. RR = raw reads (i.e. unfiltered reads) A. Assemblies performed with SPADES B. Assemblies performed with A5-MISEQ.

2.4.2.4 Optimizing the assembly of *Enterocytozoon hepatopenaei* by assessing read coverage

A filtering approach was applied at the assembled contig level in order to remove likely contamination from this assembly. In this filtering approach the unfiltered Illumina paired-end reads were mapped back onto the A5-MISEQ raw read assembly (Figure 2.5). 99.76 % of Illumina reads mapped back onto the assembled contigs. However, the coverage distribution of the reads across the assembled contigs displayed a bimodal distribution with almost half the total length of the assembly presenting a coverage below 37 X (Figure 2.5). Upon querying contigs with low coverage (< 37 X) against the NCBI nucleotide database, 2236 contigs appeared to have a bacterial origin and accounted for 2.2 Mbp of the assembly. Most of these bacterial hits had high (~99 %) identity to sequences belonging to *Acinetobacter* sp. Also, these contigs had a GC-content percentage averaging at 35.2 %. Amongst these low coverage contigs, only one had a microsporidian BLAST hit which represented a fragment of rDNA (Figure 2.5 left).

Upon querying contigs with high coverage (> 37 X) against the NCBI database, only 23 contigs (0.06 Mbp) were predicted to have a bacterial origin, 298 (0.36 Mbp) were not assigned to any taxa, 19 (0.03 Mbp) were assigned to other eukaryotic lineages other than Microsporidia and 287 (2.44 Mbp) were predicted to have a microsporidian origin. The GC-content of the predicted microsporidian contigs averaged at 26.65 %. This value was used to retrieve likely microsporidian contigs that were not assigned to any taxa by the BLASTP/N analysis (Figure 2.5 Purple). That is, contigs not assigned to any taxa by the BLASTP/N analysis but had an average GC-content below 27 % were retained as microsporidian whereas contigs that possessed GC-content above 27 % were discarded as contamination. Only contigs with a minimum length of 501 bp were retained. In total, 538 microsporidian contigs (2.78 Mbp) were predicted from these analyses. The N50 value of this assembly was 20738 (Table 2.9).

PACBIO subreads for the *E. hepatopenaei* genome, assembled *de novo* with CANU (v1.3), a forked version of the CELARA assembler (Myers et al. 2000) was

supplied by collaborators from Mahidol University, Thailand. The filtered Illumina reads were subsequently used to perform erroneous base call correction on the assembled PACBIO reads with the genome assembly improvement tool, PILON (v1.18) (Walker et al. 2014). A consensus assembly between the filtered Illumina assembly described above and corrected PACBIO reads together with contigs found in only one of these two assemblies were retrieved. Contaminating bacterial sequences that persisted in the assembly were removed using a combination of both BLASTP and BLASTX searches. Statistics of the final assembly are summarised in Table 2.16

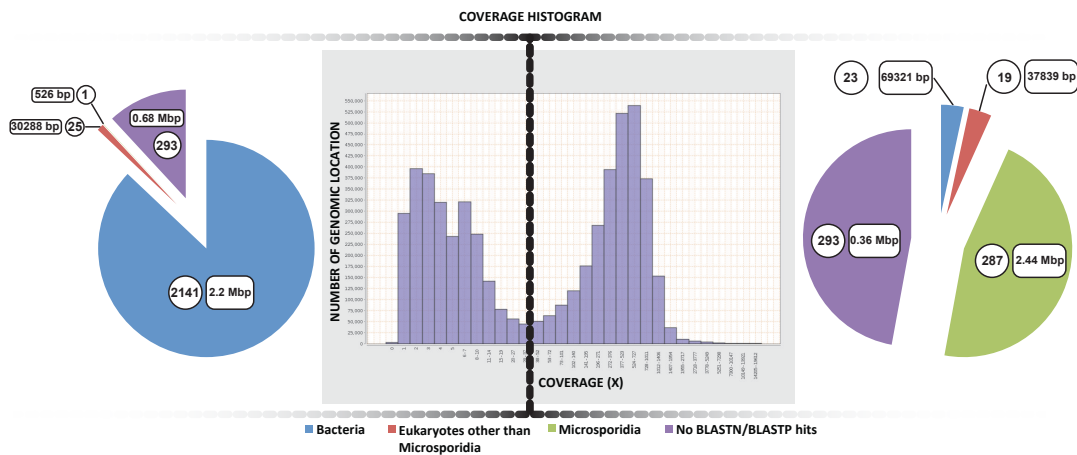


Figure 2.5 Optimizing assembly of *Enteroctozoon hepatopenaei*. Centre: Unfiltered Illumina paired-end read coverage distribution on A5-MISEQ-assembled contigs of *E. hepatopenaei* displaying a bimodal distribution. Left: BLASTP/N taxonomic assignment of contigs with an average read coverage below 37x. Right: BLASTP/N taxonomic assignment of contigs with an average read coverage above 37x. Numbers in circles and curved rectangles stand for number and lengths of assembled contigs respectively assigned to a particular taxon.

Table 2.9: Statistics for the final assembly of *Enteroctozoon hepatopenaei*

Assembly Statistics	A5
# Contigs (>= 0 bp)	538
# Contigs (>= 1000 bp)	302
Total length (>= 0 bp)	2782625
Total length (>= 1000 bp)	2613694
Largest contig	118107
GC (%)	25.85
N50	20738
# N's per 100 kbp	73.06

2.4.3 Preliminary assembly of the *Enterospora canceri* genome

Three sequencing runs were performed on the genome of *Enterospora canceri*. The first two sequencing attempts (RUN1 and 2) were performed on the same DNA sample where as the third attempt (RUN3) was performed on a new DNA

sample extracted on a later date. RUN1 resulted in a total of 2,049,868 paired-end Illumina reads, with sequence length of 251 bp. FASTQC-informed-manual trimming of poor quality sequences (partial adapter sequences and poor quality 3' ends) resulted in a reduction of read length to 70 bp. RUN2 resulted in a total of 7,960,664 paired-end Illumina reads, with sequence length ranging between 250-252 bp. Trimming of remnant adapter sequences and poor quality 3'-ends resulted in a read length reduction to 131 bp. RUN3 resulted in a total of 4,632,816 paired-end Illumina reads, with a sequence length of 301 bp. Here, FASTQC results did not prompt the need for further trimming. In all three runs, the final reads had mean quality score above 30. The preliminary assemblies performed with SPADES and A5-MISEQ on the Illumina reads from RUN1 were 10.98 Mbp and 8.65 Mbp in length respectively with the A5-MISEQ assembly presenting the highest N50 value of 1820 (Table 2.10). The Illumina reads from RUN2 resulted in smaller assembly sizes for both assembly programs. The SPADES assembly resulted in an assembly size of 3.02 Mbp whereas the A5-MISEQ assembly yielded an assembly size of 2.73 Mbp (Table 2.11). Here, the SPADES and A5-MISEQ assembly programs attained almost identical N50 values, 23096 and 23048 respectively. The Illumina reads from RUN3 resulted in assembly sizes of 4.35 and 6.75 Mbp for the A5_MISEQ and SPADES assemblies respectively (Table 2.12). The A5-MISEQ assembly had an N50 value of 5865 whereas the SPADES assembly an N50 value of 1248. Considering the assemblies performed with reads from all three RUNs, RUN1 resulted in the largest assembly size suggesting it was the most likely to contain contamination. Upon aligning the Illumina reads from RUN1 onto its respective SPADES assembly, only 53.95 % of the reads mapped and they displayed a unimodal coverage distribution (Figure 2.6) unlike the bimodal distribution displayed by the *E. hepatopenaei* assembly. Consequently, elimination of contaminant contigs by assessing coverage distribution, as performed in Section 2.4.2.4, could not be performed here.

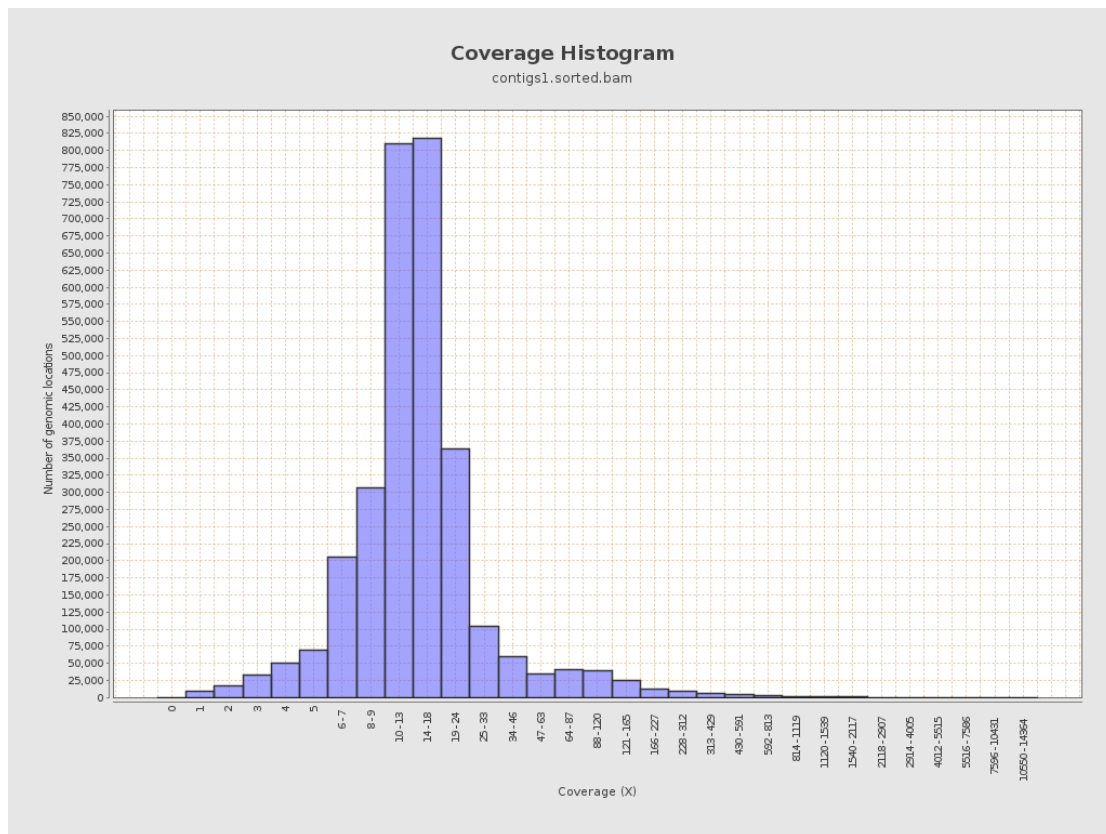


Figure 2.6: Coverage distribution of *Enterospora canceri*'s Illumina reads on its SPADES-assembled genome from RUN1 displaying a unimodal distribution.

Table 2.10: Comparing SPADES and A5-MISEQ assemblies performed on the raw Illumina reads of *Enterospora canceri* obtained from the first sequencing run. Only contigs > 500 bp were used.

RUN1	SPADES	A5-MISEQ
# Contigs (>= 0 bp)	9034	6278
# Contigs (>= 1000 bp)	1955	1798
Total length (>= 0 bp)	10974927	8644614
Total length (>= 1000 bp)	6475292	5688959
Largest contig	250209	177371
GC (%)	41.40	41.22
N50	1441	1820
# N's per 100 kbp	0.00	1353.33

Table 2.11: Comparing SPADES and A5-MISEQ assemblies performed on the raw Illumina reads of *Enterospora canceri* obtained from the second sequencing run. Only contigs >500 bp analysed.

RUN2	SPADES	A5-MISEQ
# Contigs (>= 0 bp)	661	461
# Contigs (>= 1000 bp)	354	304
Total length (>= 0 bp)	3020677	2727163
Total length (>= 1000 bp)	2811399	2615959
Largest contig	140170	89024
GC (%)	40.35	40.25
N50	23096	23048
# N's per 100 kbp	1.52	420.03

Table 2.12: Comparing SPADES and A5-MISEQ assemblies performed on the raw Illumina reads of *Enterospora canceri* obtained from the third sequencing run. Only contigs > 500 bp analysed.

RUN3	SPADES	A5-MISEQ
# Contigs (>= 0 bp)	5892	1897
# Contigs (>= 1000 bp)	872	835
Total length (>= 0 bp)	6713369	4345220
Total length (>= 1000 bp)	3649838	1883728
Largest contig	94614	130884
GC (%)	39.74	40.39
N50	1248	5865
# N's per 100 kbp	0.00	498.87*

2.4.4 Assembly optimization of the genome of *Enterospora canceri*

A preliminary Illumina read GC filtering on the assemblies performed with reads from all three RUNs revealed that the maximum number of 381-core microsporidian protein set was only attained with the unfiltered read assembly and not with GC-filtered read assemblies (Figure 2.7). Secondly initial analysis performed on a subset of identified microsporidian contigs revealed that they possessed a GC composition higher than 40%. As this initial data suggested that the *Ent. canceri* genome could potentially be of high GC composition, a filtering approach at the contig level that did not rely on GC content was employed to detect and remove likely contamination.

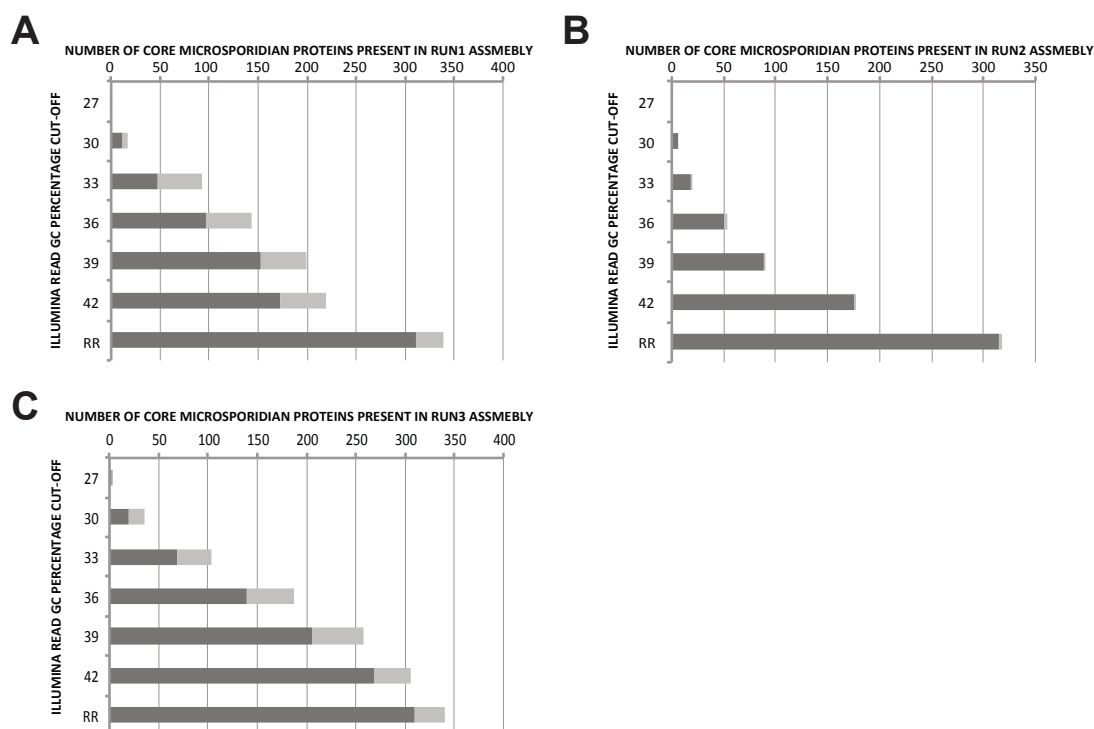


Figure 2.7: Estimating completeness of assembled genome of *Enterospora canceri* at different GC content cut-offs. Deep grey = KRONA predicted microsporidian proteins. Light grey = proteins predicted as non-

microsporidian or not assigned to any taxa by KRONA. RR = raw reads (i.e. non-filtered reads). A, B, and C, represent assemblies for first second and third sequencing RUNs respectively.

2.4.4.1 Establishing consensus assembly between RUN1 and RUN2

Contigs from RUN1 and 2 assemblies were aligned with the NUCMER program (Kurtz et al. 2004) incorporated within the QUASt (Gurevich et al. 2013) package in order to identify true microsporidian contigs. A total length of 3.67 Mbp from RUN1 aligned to 3.0 Mbp from RUN2 and these contigs were retained. The unaligned assembly from both RUNs was discarded as contamination. In order to identify microsporidian contigs present in the discarded, unaligned contigs and non-microsporidian contigs present in both runs that aligned to each other, further assembly optimisation steps were employed.

2.4.4.1.1 Filtering aligned RUN1 assembly by k-mer coverage, BLAST and contig length

Contigs from RUN1 that aligned to contigs from RUN2 were further analysed to ascertain their microsporidian origin. These contigs could be grouped into three distinct subsets: Contigs with very high (> 5), medium (~ 3) and low k-mer coverage (< 1.6) (Figure 2.8). When the lengths of the respective contigs were superimposed upon the k-mer coverage map, contigs with high and low coverage were observed to be relatively short (< 15 Kbp) as compared to contigs with medium coverage (> 15 Kbp) (Figure 2.8).

Upon querying a subset of ORFs from these contigs and/or their entire nucleotide sequence against the NCBI database, high-coverage-short-length, medium-coverage-short-length and low-coverage-short-length contigs often resulted in strong arthropod BLAST hits or weak BLAST hit for other eukaryotic lineages other than Fungi or Arthropoda (Figure 2.8). Arthropod BLAST hits often belonged to the Brown Crab, *Seylla olivacea*. All examined contigs that were longer than 15 Kbp resulted in microsporidian BLAST hits (Figure 2.8). In light of these results, contigs with k-mer coverage values ranging between 1.6 and 3.8, and length above 11158 bp were retained as microsporidian contigs (Figure 2.8). They consisted of 42 contigs, which totalled to 2.20 Mbp.

In order to retrieve microsporidian contigs that may have been discarded during the alignments of the two RUNs (Section 2.4.4.1) and/or the above mentioned coverage/filtering process, the 8992 discarded contigs (8.78 Mbp) from RUN1 were queried against the NCBI protein/nucleotide database: 2866 contigs (2.66

Mbp) had non-microsporidian eukaryotic BLAST hits, 2339 (3.45 Mbp) contigs had bacterial BLAST hits and only 46 contigs (0.18 Mbp) had microsporidian BLAST hits (these were added to the list of 42 microsporidian contigs). The remaining 3741 contigs (2.49 Mbp) were not assigned to any taxa by the KRONA tool. In summary, this filtering process identified 88 contigs to be of microsporidian origin. In order to assess if the filtering process had affected the completeness of the assembled genome negatively, ORFs from the unfiltered and the filtered assembly were queried against a set of 381-core microsporidian protein-set with BLASTP. The filtered and unfiltered assembly contained 340 and 339 proteins respectively out of 381 suggesting minimal loss in assembly completeness.

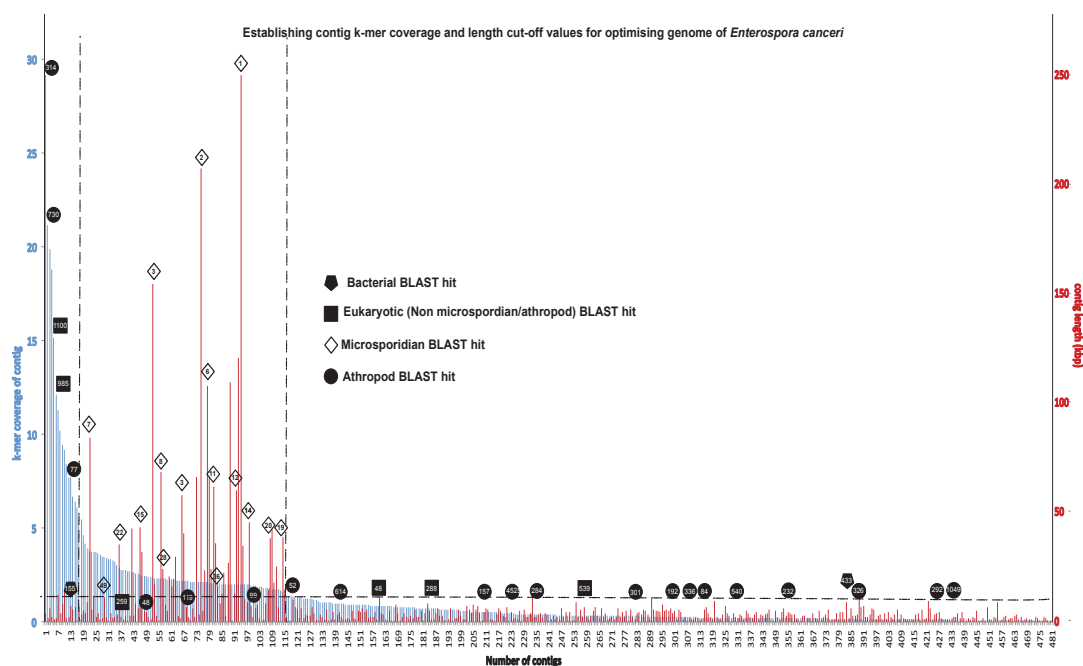


Figure 2.8: Establishing filtering parameters (dotted lines) for the genome of *Enterospora canceri*. K-mer coverage of *Enterospora canceri* SPADES-assembled contigs from RUN1 that aligned to RUN2 contigs (Blue bars). Superimposed red bars are the lengths in bp for corresponding contigs. Numbers within shapes are the contig ID numbers.

2.4.4.1.2 Reassembling RUN1 with reads that mapped onto filtered contigs

The average k-mer and read coverage of the 88 filtered contigs from RUN1 was 1.9 and 16.1 X respectively. In order to increase these values and ultimately optimize the entire assembly, reads from RUN1 were remapped back onto the 88 filtered contigs. 26.98 % of the total number of reads mapped back onto the filtered contigs (Table 2.13). The mapped reads were used to reassemble the *Ent. canceri* genome with the SPADES pipeline. This resulted in a total assembly size of 2.31 Mbp (contigs > 500 bp) and an N50 value of 25424. Interestingly, this new assembly was further fragmented into 219 contigs (Table 2.13).

Furthermore, its average k-mer and read coverage had increased, albeit minimally to 12.3 and 16.3 X respectively.

Table 2.13: Statistics for the genome of *Enterospora canceri* assembled with reads that remapped onto filtered contigs.

STATISTICS	RUN1	RUN2
# Contigs (>= 0 bp)	219	579
# Contigs (>= 1000 bp)	175	348
Total length (>= 0 bp)	2309440	3325067
Total length (>= 1000 bp)	2277950	2848388
Largest contig	82875	139989
GC (%)	40.66	40.55
N50	25424	32978
# N's per 100 kbp	0.69	0.00

2.4.4.1.3 Reassembling RUN2 with reads that mapped onto filtered contigs

625 (2.30 Mbp) out of the initial 661 (3.02 Mbp) contigs from RUN2 aligned to RUN1's filtered 88 contigs. Prior to the remapping of reads onto the 625 contigs, a preliminary analysis to assess whether the above-described filtering process had affected assembly completeness was performed by querying a set of 381-core microsporidian proteins with BLASTP against ORFs encoded by the 625 contigs. This showed that both the filtered and unfiltered contigs contained 340 out of the 381 queried proteins, suggesting the filtering process did not affect genome completeness. These 625 contigs were extracted and their corresponding trimmed reads (reads from RUN2) were subsequently remapped onto them. 57.33 % of the total number of reads mapped back onto the filtered contigs (Table 2.14). These reads were used to reassemble the genome of *Ent. canceri*. This resulted in a total assembly size of 3.33 Mbp and an N50 value of 32978 (Table 2.13). Interestingly, the overall k-mer coverage increased from 5.8 to 127 whereas the read coverage reduced from 146.2 to 134.9 X.

2.4.4.1.4 De novo assembly of genome of *Enterospora canceri* by combining reads from RUN1 and RUN2

Reads that remapped to filtered RUN1 and RUN2 assembly were used here. For this, the multiple library option in the SPADES assembly program was invoked to utilise reads from both RUNs to perform a single assembly. After the removal of short contigs (< 500 bp), the final assembly was 2.27 Mbp long and consisted of 167 contigs (Table 2.15). The average k-mer and read coverage was 120.46 and

65.5X respectively. In order to assess level of contamination in the final assembly, predicted ORFs were queried against the NCBI database and the results parsed by the KRONA taxonomic profiling tool. This showed that out of the 167 assembled contigs, 109 were microsporidian, 1 was bacterial and the remaining 57 had an unknown origin. Furthermore, an assessment of genome completeness showed that out of 381-core microsporidian proteins, 309 microsporidian proteins were found in the assembly (Figure 2.9).

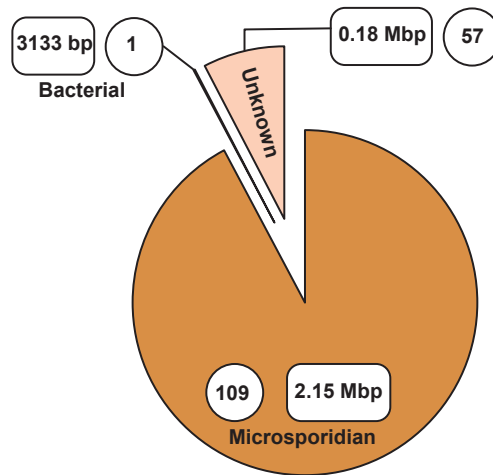


Figure 2.9: Assessing levels of contamination in genome of *Enterospora canceri* reassembled from RUN1 and RUN2 Illumina reads.

Table 2.14: Remapping statistics RUN1 and 2

STATISTICS	RUN1 ASSEMBLY	RUN2 ASSEMBLY
# READS (TOTAL)	2041260	7026231
# MAPPED READS	550811 (26.98 %)	4028436 (57.33 %)

Table 2.15: *De novo* assembly statistics of *Enterospora canceri* genome assembled with reads from RUN1 and RUN2

statistics	SPADES
# contigs (>= 0 bp)	167
# contigs (>= 1000 bp)	136
Total length (>= 0 bp)	2266597
Total length (>= 1000 bp)	2245549
Largest contig	218090
GC (%)	40.70
N50	37212
# N's per 100 kbp	0.04

2.4.4.1.5 Filtering RUN3 Assembly by k-mer coverage

This was performed by ordering contigs from the initial RUN3 assembly (Table 2.12) by decreasing k-mer size and observing their corresponding length distribution (Figure 2.10). Contigs could be grouped into three distinct subsets:

Contigs with very high (> 600), medium (~500) and low k-mer coverage (< 130) (Figure 2.10).

Upon querying ORFs on these contigs and/or their entire nucleotide sequence against the NCBI database, it was clear that microsporidian contigs were present in all k-mer coverage regions (Figure 2.10) and not clustered in only low and high k-mer regions as with RUN1 and 2. Since k-mer coverage was unsuccessful in eliminating contaminant contigs, filtering was solely based on BLAST results. Here, contigs that had positive BLAST hits for arthropod and bacteria sequences were removed from the assembly. At this point, only 1300 (3.52 Mbp) out of the initial 1897 contigs remained. In order to eliminate misassemblies caused by contaminant reads, the filtered Illumina reads were mapped back onto the 1300 filtered contigs and the mapped reads were used to reassemble the genome. The read coverage of the reassembled genome was 329 X. The assembly statistics of the reassembled genome have been summarised in Table 2.16.

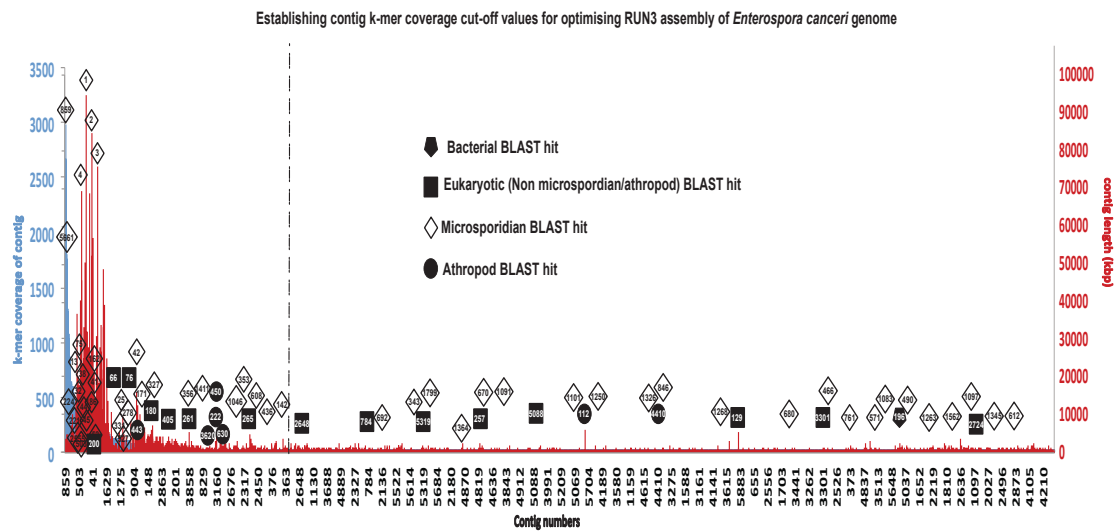


Figure 2.10: Establishing filtering parameters for the RUN3 assembly of *Enterospora canceri*'s genome: Comparing K-mer coverage distribution (Blue bars) and contig length (Red bars) across SPADES assembly.

To assess if the filtering process had affected the completeness of the reassembled genome negatively, ORFs from the filtered assembly were queried against a set of 381-core microsporidian protein-set with BLASTP. Both the filtered and unfiltered assembly contained 340 proteins suggesting no loss in assembly completeness.

Table 2.16: *De novo* assembly statistics of *Enterospora canceri* genome assembled with Illumina reads that mapped onto the filtered RUN3 contigs

STATISTICS	
# Contigs (>= 0 bp)	1416
# Contigs (>= 1000 bp)	620
Total length (>= 0 bp)	3708854
Total length (>= 1000 bp)	3095845
Largest contig	130884
GC (%)	40.09
N50	10121
# N's per 100 kbp	0.00

2.4.4.1.6 Assembling the genome of *Enterospora canceri* with reads from RUN1, 2 and 3

Only reads that remapped onto filtered contigs from RUN1, 2 and 3 were used here. Statistics of this assembly are summarized in Table 2.17. Interestingly, this assembly was more fragmented with a total assembly size of 2.3 Mbp (Smaller than some of the individual assemblies).

Table 2.17: *De novo* assembly statistics of *Enterospora canceri* genome assembled with reads from RUN1, 2 and 3

Final <i>Enterospora canceri</i> statistics	
# Contigs (>= 0 bp)	1635
# Contigs (>= 1000 bp)	856
Total length (>= 0 bp)	2286859
Total length (>= 1000 bp)	1743776
Largest contig	9089
GC (%)	40.29
N50	1762
# N's per 100 kbp	0.09

2.4.4.2 The final *Enterospora canceri* assembly

In the end, the filtered assembly performed with reads from RUN3 was submitted and not the assembly resulting from the combination of reads from the three runs. Following further BLAST-based filtering processes, elimination of low read coverage contigs and GENBANK filtering processes, the final assembly size was 3.1 Mbp spanning across 537 contigs (see Table 2.18 for statistics).

Table 2.18: Statistics for assembled genomes submitted to GENBANK

statistics	<i>Hepatospora eriocheir</i>	<i>Hepatospora eriocheir canceri</i>	<i>Enterospora canceri</i>	<i>Enterocytozoon hepatopenaei</i>
Assembly size (Mb)[Jellyfish estimate]	4.66	4.84	3.10	3.26
Contigs	1300	2344	537	64
Mean coverage (X)	4477.89	63.18	288	363
Contig N50 (bp)	17583	3349	11128	125008
GC content (%)	22.44	23.16	40.15	25.45
GC content coding (%)	25.58	25.41	41.95	27.81
Coding regions (%)	42.39	40.31	59.50	71.95
Splicing machinery	8/29 genes	8/29 genes	8/29 genes	6/29 genes
Genes	2716	3058	2179	2540

2.4.5 Predicted open reading frames (ORFs) in the sequenced genomes

The MAKER annotation pipeline predicted a total of 2673, 2933, 1859 and 2081 ORFs for the genomes of *H. eriocheir*, *H. eriocheir canceri*, *Ent. canceri* and *E. hepatopenaei* respectively. After further manual editing and identification of ORFs missed by MAKER, the final number submitted to GENBANK were 2716, 3058 and 2178 for the genomes of *H. eriocheir*, *H. eriocheir canceri* and *Ent. canceri* respectively (Table 2.18). After the merge of MISEQ Illumina data with PACBIO

scaffolds provided by collaborators in Thailand for *E. hepatopenaei*, the final number of predicted ORFs submitted to GENBANK was 2540 (Table 2.18).

2.4.6 Conserved motifs upstream of start codons

Analysis of the 100 bp region upstream of all predicted ORFs in the sequenced genomes revealed an enrichment of GGGTAAAA nucleotide sequence, which often occurred 25 bp upstream of start codons. This motif was ubiquitous across all Enterocytozoonidae sequenced genomes. The remaining motifs were however species specific with TTTTATT highly represented in the genomes of the *Hepatospora* spp. and AACAA and AAAATA highly represented in the genomes of *Ent. canceri* and *E. hepatopenaei* (Figure 2.11).

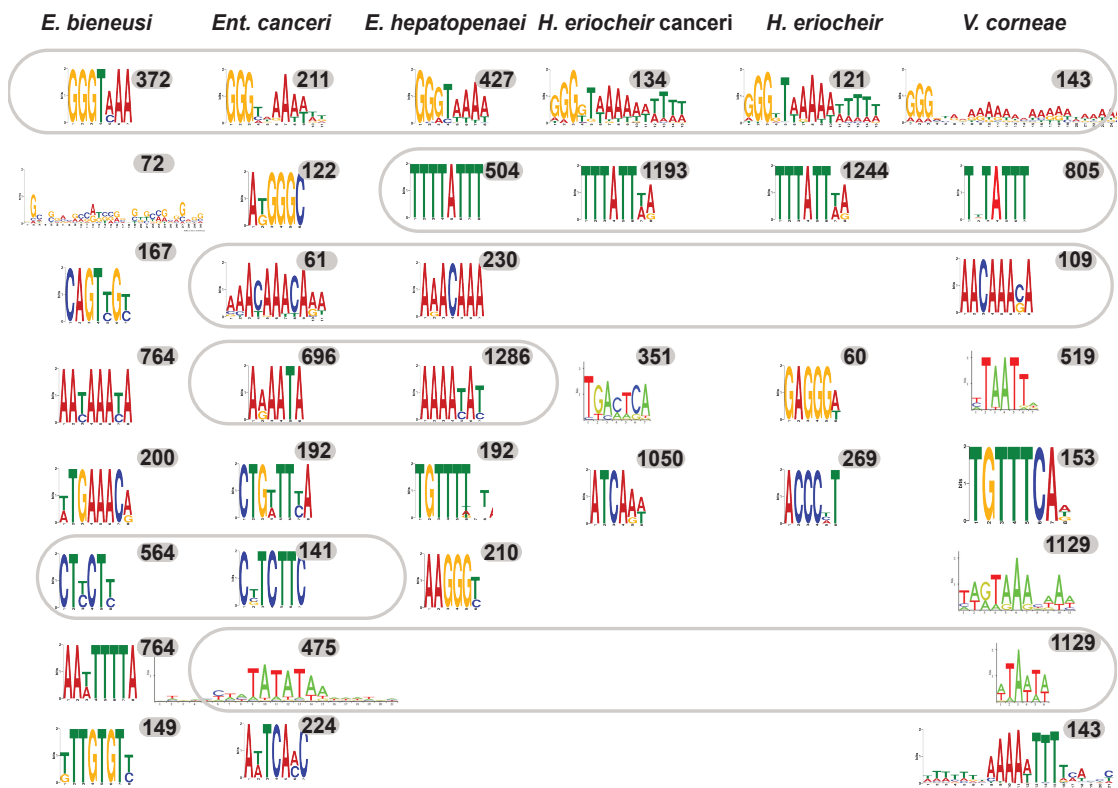


Figure 2.11: Comparing putative regulatory motifs upstream of microsporidian ORFs. Numbers on grey backgrounds represent number of occurrences of its corresponding motif in that particular genome. Sequences shared between different lineages have been grouped with grey oblong boxes.

2.4.7 tRNAs found in the sequenced genomes and frequencies of their corresponding amino acids

The genomes of all sequenced microsporidians in this study encoded tRNAs for all 20 amino acids. Each genome often encoded a single copy of each tRNA (Figure 2.12). Some synonymous tRNAs were absent from either one or all analysed genomes. Those absent from all sequenced genomes were: CGG-Ala,

UUA-Asn, CUA-Asp, ACA-Cys, CCA/CCC-Gly, GUA-His, UAG-Ile, GAG-Leu, GGG/GGC-Pro, AGG/UCA-Ser, AUA-Tyr, CAG-Val and UGG-Thr (Figure 2.12). tRNA numbers for the genome of *E. bieneusi* showed a relationship with amino acid usage (See Figure 2.12 and 2.13 for Ile, Lys, Leu and Asn). Other taxa analysed in this study did not display such strong relationship between their tRNA abundance and amino acid usage.

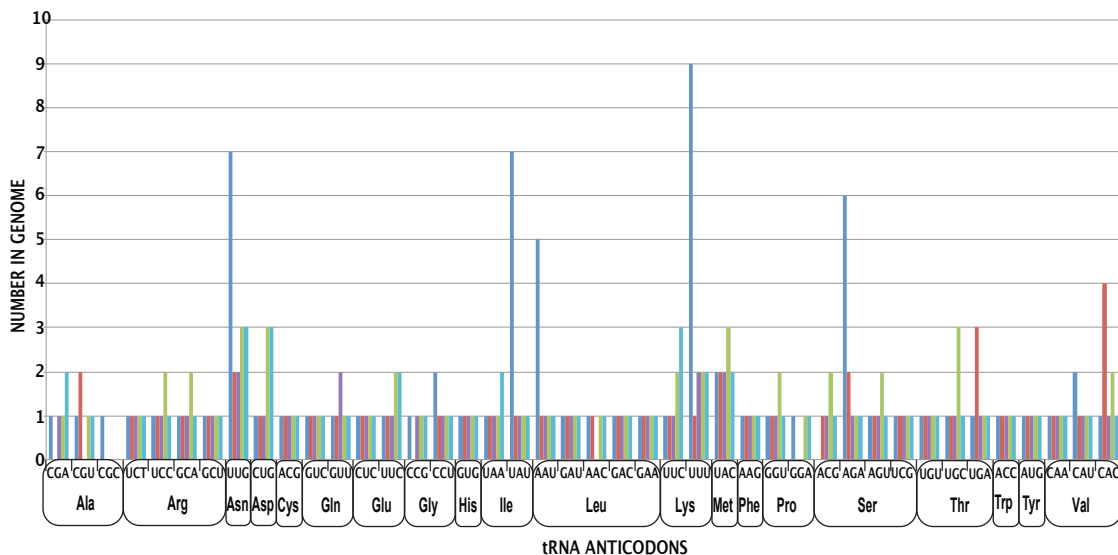


Figure 2.12: tRNAs in the Enterocytozoonidae: *Enterocytozoon bieneusi* (blue), *Enterospira canceri* (red), *Enterocytozoon hepatopenaei* (violet), *Hepatospora eriocheir canceri* (green), *Hepatospora eriocheir* (light blue). Synonymous anticodons grouped by round-edged squares.

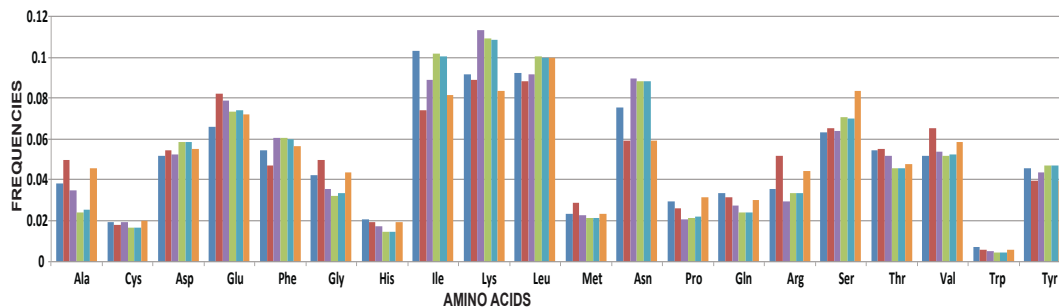


Figure 2.13: Frequency of each amino acid in predicted ORFs of the Enterocytozoonidae and *Vittaforma*. *Enterocytozoon bieneusi* (blue), *Enterospira canceri* (red), *Enterocytozoon hepatopenaei* (violet), *Hepatospora eriocheir canceri* (green), *Hepatospora eriocheir* (light blue) and *Vittaforma. corneae* (orange).

2.4.8 GC content and synonymous codon usage bias in the Enterocytozoonidae

Hepatospora species and *E. hepatopenaei* genomes contained an average GC content below 26 % whereas that of *Ent. canceri* was 40 % (Table 2.18). The GC content of the ORFs of all four newly sequenced members of the Enterocytozoonidae family was only slightly higher than the overall genomic GC content (Table 2.18). ORFs of *Hepatospora* spp. and *E. hepatopenaei* presented with higher GC content at the first and second codon positions as compared to

the GC content at their third codon position (Figure 2.14). *Ent. canceri* had slightly higher GC content at its first and second codon positions as compared to *Hepatospora* spp. and *E. hepatopenaei* (Figure 2.14). In contrast with observations in the genomes of *Hepatospora* spp. and *E. hepatopenaei*, the third codon position of *Ent. canceri* ORFs had a higher GC content as compared to first and second codon positions (Figure 2.14).

The average synonymous codon usage order (SCUO) for ORFs in each of the genomes sequenced in this study together with that of *E. bieneusi* (publicly available) were analysed. SCUO is a measure of bias that ORFs within a genome display towards specific synonymous codons with SCUO=0 representing no bias and SCUO=1 representing highest bias. ORFs from *E. hepatopenaei* had a SCUO value of 0.34. 180 (8.65 %) ORFs from this genome displayed SCUO values above 0.5. ORFs from both *Hepatospora* spp. featured SCUO values of 0.40. 372 (14.03 %) and 377 (14.21 %) ORFs from *H. eriocheir canceri* and *H. eriocheir* had a SCUO value above 0.5. In contrast, only 16 ORFs (0.86 %) from *Ent. canceri* presented a SCUO value above 0.5 and the average SCUO value for all its ORFs was 0.20. A Wilcoxon Two Sample Test (WTST) performed on the ORFs of the two *Hepatospora* strains showed that there was no difference in their SCUO distribution (p-value=0.944). The same test performed on the gene set of *E. hepatopenaei* against *H. eriocheir canceri* and *H. eriocheir* produced p-values of 1.095e-95 and 4.412e-94 respectively, suggesting a considerable difference in SCUO distribution between *E. hepatopenaei* and *Hepatospora* spp. The SCUO distribution comparisons between the ORFs of *Ent. canceri* and the *Hepatospora* strains showed the greatest disparity (p-value 0). A p-value of 4.964e-300 was recorded for the SCUO comparisons between *Ent. canceri* and *E. hepatopenaei* (Figure 2.14).

With regards to frequency at which synonymous codons were used in the analysed genomes, codons with a G or C in their third position were present at higher levels in ORFs belonging to *Ent. canceri* compared to other analysed taxa (Figure 2.15).

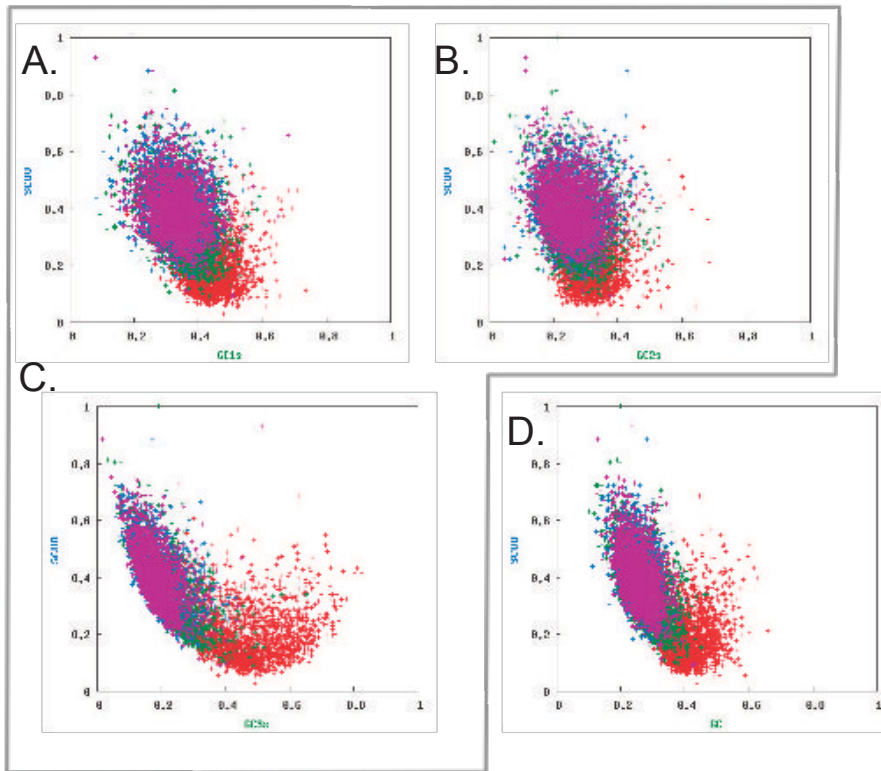


Figure 2.14: Visualization of synonymous codon usage bias (SCUO) against GC distribution for ORFs in the genomes of the Enterocytozoonidae. *Enterocytozoon bienensei* (red), *Hepatospora eriocheir* (purple), *Hepatospora eriocheir* (red) and *Enterocytozoon hepatopenaei* (green). Analysis for A, first, B, second and C, third codon positions are labelled GC1, 2, 3 respectively. An average of values in graph A, B and C is computed in Graph D.

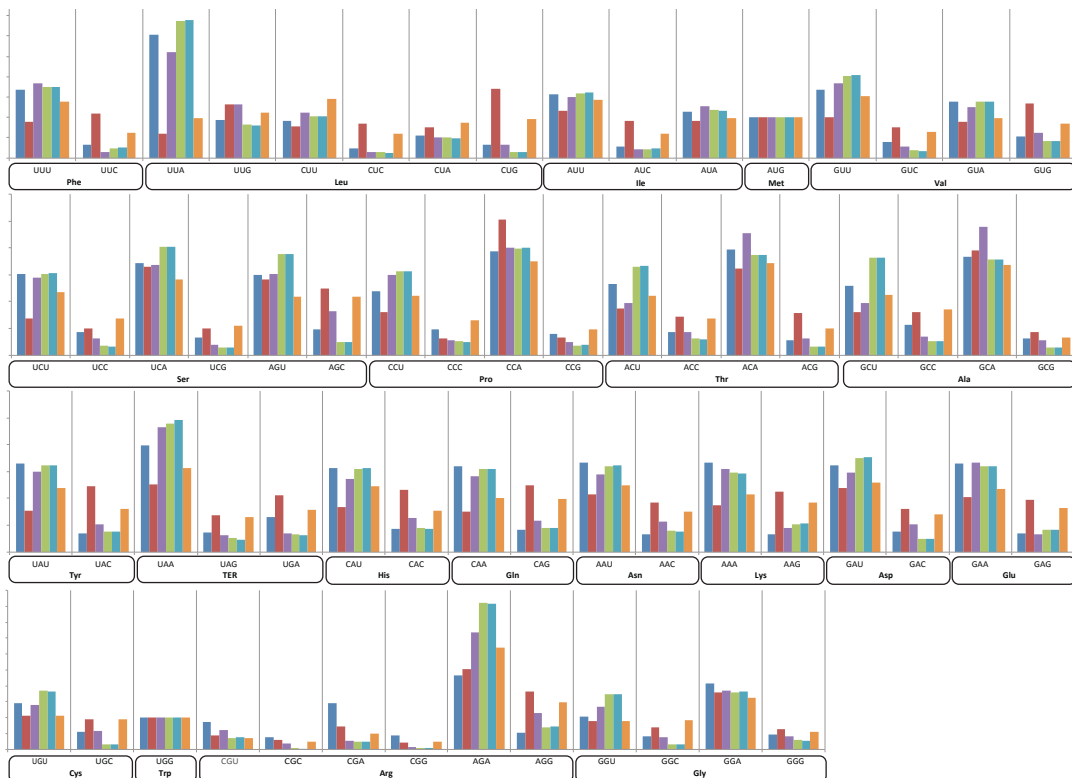


Figure 2.15: Comparing codon usage frequencies between *Enterocytozoon bieneusi* (blue), *Enterospora canceri* (red), *Enterocytozoon hepatopenaei* (violet), *Hepatospora eriocheir canceri* (green), *Hepatospora eriocheir* (light blue) and *Vittaforma corneae* (orange). A total of 1500 ORFs was used for all species.

2.4.9 Identifying transposable elements and repetitive DNA sequences

2.4.9.1 DFAM predictions

The majority of retrotransposons identified in the genome of *Ent. canceri* belonged to the Gypsy or CR1 superfamilies whereas transposons belonged to the Merlin or TcMar-Tc1 superfamilies. A total of 53227 bp of *Ent. canceri*'s genome was predicted to encode transposable elements. For the genomes of *E. hepatopenaei* and the *Hepatospora* spp., almost all retrotransposons belonged to the Gypsy superfamily whereas the transposons belonged to the TcMar-Tc1 superfamily. The total length of transposable elements in *H. eriocheir canceri* was 113655 bp whereas in *H. eriocheir* that was 65541 bp and *E. hepatopenaei* was 10438 bp. The DFAM supplementary table (Appendix 9) contains more details on the coordinates of the identified repetitive DNA sequences.

2.4.9.2 REPEATFINDER predictions

Retrotransposons predicted for the *Ent. canceri* genome predominantly belonged to the CR1 superfamily whereas all transposons belonged to the hAT-Charlie superfamily. A total of 12969 bp was predicted to encode transposable elements from this genome. The genome of *E. hepatopenaei* was predicted to contain a total of 1043 bp of transposable elements. All of which were retrotransposons or unclassified. The total length of transposable elements in *H. eriocheir canceri* was 606 bp whereas that of *H. eriocheir* was 995 bp. All transposons in the *Hepatospora* spp. belonged to the hAT-Charlie superfamily whereas retrotransposons belonged to the CR1, SINE and LINE family. Upon close inspection, some transposable elements identified for the analysed genomes were genomic regions that had high sequence similarity to a short fragment of a queried transposable element. That is, whereas the length of a transposable element is characteristically in the range of 1 to 5 kbp (Muñoz-López & García-Pérez 2010), the genomic length identified by REPEATFINDER for some of the analysed genomes was occasionally as short as 45 bp.

2.4.10 Phylogenomics of the Microsporidia

With *Mitosporidium daphniae* used as an outgroup to the microsporidian lineage, *Ent. canceri* branched as the closest relative to *E. bieneusi* and *Hepatospora* spp. branched as the most basal group of the *Enterocytozoonidae* clade. Both maximum likelihood and Bayesian inference analyses produced a phylogenetic tree with a similar topology (Figure 2.16).

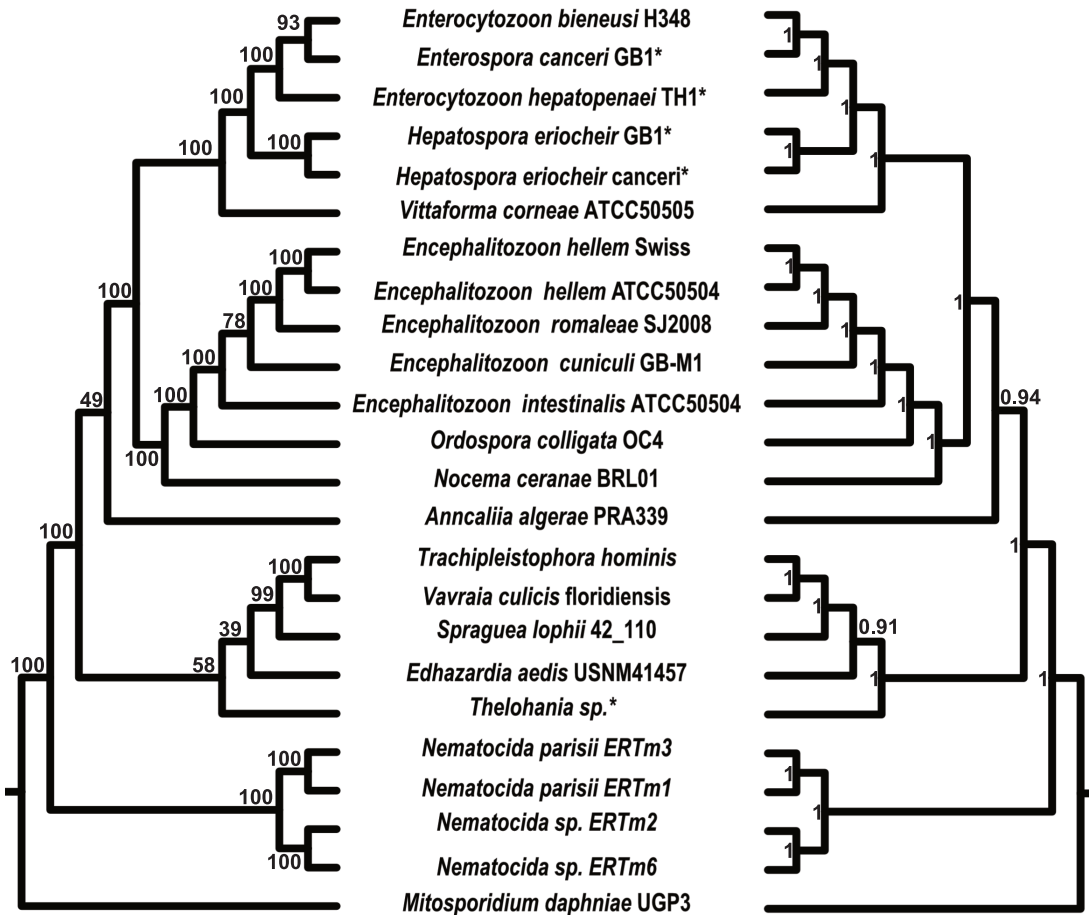


Figure 2.16: Cladogram of 23 microsporidian species. Left: Best-scoring maximum likelihood tree out of 100 bootstrapped trees. Values on nodes represent bootstrap support. Right: Bayesian analysis. Numbers on nodes represent posterior probability values. Analyses were performed on a concatenated alignment of 21 proteins. Species with * are those whose genomic data were produced in this study.

2.4.11 Mapping of metabolic functions onto the microsporidian phylum

A search for the presence of key metabolic pathways in the above-mentioned genomes shows that the *Enterocytozoonidae* clade lacked at least a gene in each metabolic pathway analysed (Figure 2.17). Independent losses of genes in other microsporidian lineages were also observed. Amongst the metabolic processes analysed, glycolysis and trehalose metabolism were the only pathways for which all microsporidians with exception of the *Enterocytozoonidae* possessed a full gene complement.

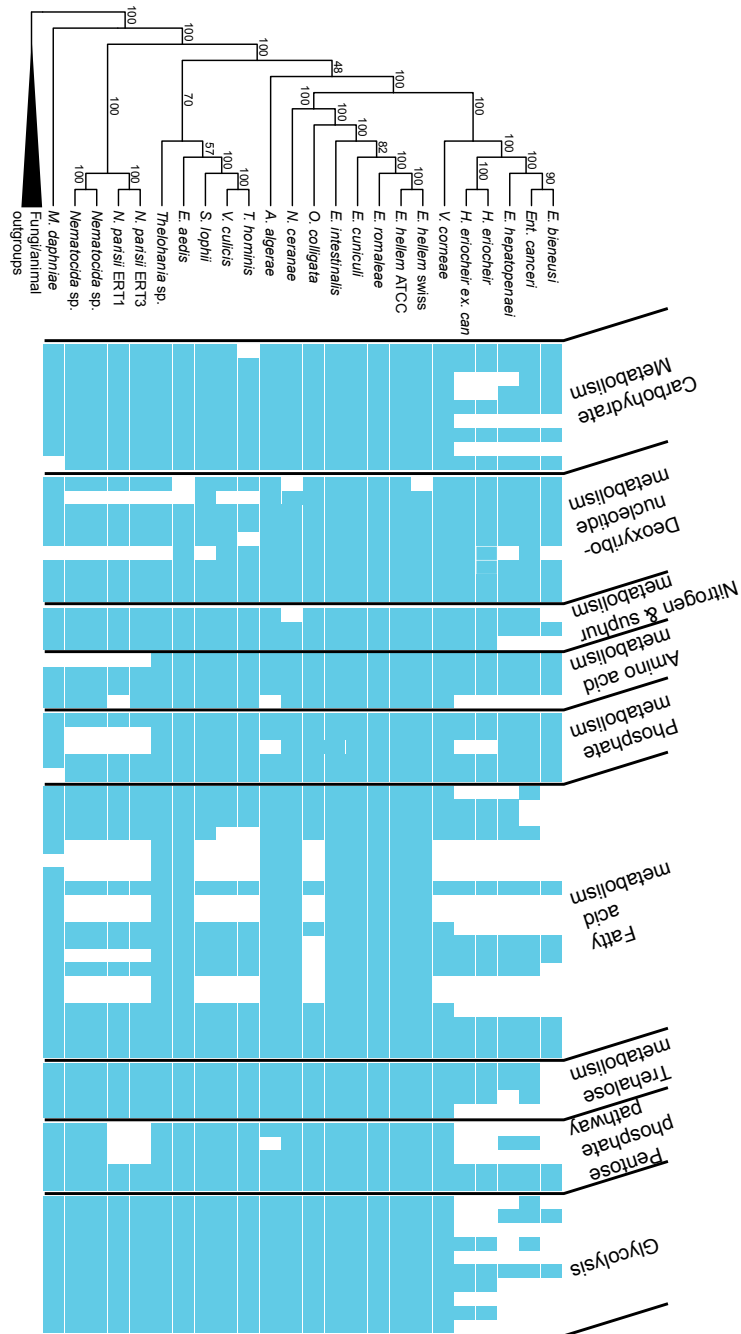


Figure 2.17: Metabolic profiling of microsporidian genomes. Filled squares represent presence of a gene involved in a metabolic pathway whereas a blank space represents absence of a gene. The phylogenetic positions in the cladogram are derived from maximum likelihood analysis performed on a concatenated alignment of 21 conserved proteins. Node values represent levels of support from 100 bootstrap replicates. Genes used in this analysis were taken from Keeling *et al.* 2010. Deoxyribonucleotide metabolism orthologs for *Encephalitozoon cuculii* proteins ECU08_0090 and ECU01_1430 were removed from analysis as these proteins are paralogous to ECU01_0180 (this has been retained). Pyruvate Dehydrogenases ECU09_1040 and ECU04_1160 grouped within glycolytic proteins in Keeling *et al.* (2010) were also removed from this analysis because they are not core glycolytic proteins. (From Wiredu-Boakye *et al.* submitted, see appendix 13)

2.4.12 Comparing the plasma membrane transporter repertoire of nuclear and cytoplasm-infecting microsporidians

Plasma membrane transporters across the sequenced microsporidian genomes were compared to those in selected publicly available microsporidian genomes. In general, genomes of members of the Enterocytozoonidae possessed fewer plasma membrane transporters with a predicted function (approximately 22) as compared to non-members of this family (approximately 47) (Figure 2.18). The V/F-type ATPases, UAA transporter, P-type ATPases, ABC, ZIP and Pho1 transporter families were ubiquitous across all genomes surveyed. The DMT, MFS, Sulp, CTL, POT, Amino acid permease, MScS, CPA1, DASS, Ca-CIC, Cation efflux, HCO₃, sugar and transmembrane amino acid transporter families were absent in at least one species (Figure 2.18) (Appendix 10). Plasma membrane transporters belonging to the HCO₃ and POT families were predicted only for non-Enterocytozoonidae species, *Trachipleistophora hominis* and *Enc. cuniculi*. The only protein family exclusive to the Enterocytozoonidae belonged to the DASS ion transporter family and this was only found in the genome of *E. bieneusi*. No transporter family with known function was predicted as exclusive to the intranuclear parasite, *Ent. canceri* (Figure 2.18) (Appendix 10). However, the genome of this intranuclear parasite harboured twice as many plasma membrane transporters with a predicted function as compared to other members of the Enterocytozoonidae family (Appendix 10).

In all analysed species, there were more predicted transporters with unknown functions than with known functions (Table 2.19) (Appendix 10). The genomes of *E. bieneusi* and *E. hepatopenaei* harboured 14 and 1 unique transporters with unknown functions respectively. The genome of *E. bieneusi* contained the highest number of single copy unique transporters with unknown functions, 10. At least 82 plasma membrane transporters with unknown functions predicted for members of the Enterocytozoonidae family also had orthologs in the genomes of *Enc. cuniculi* and/or *T. hominis* (Table 2.19) (Appendix 10).

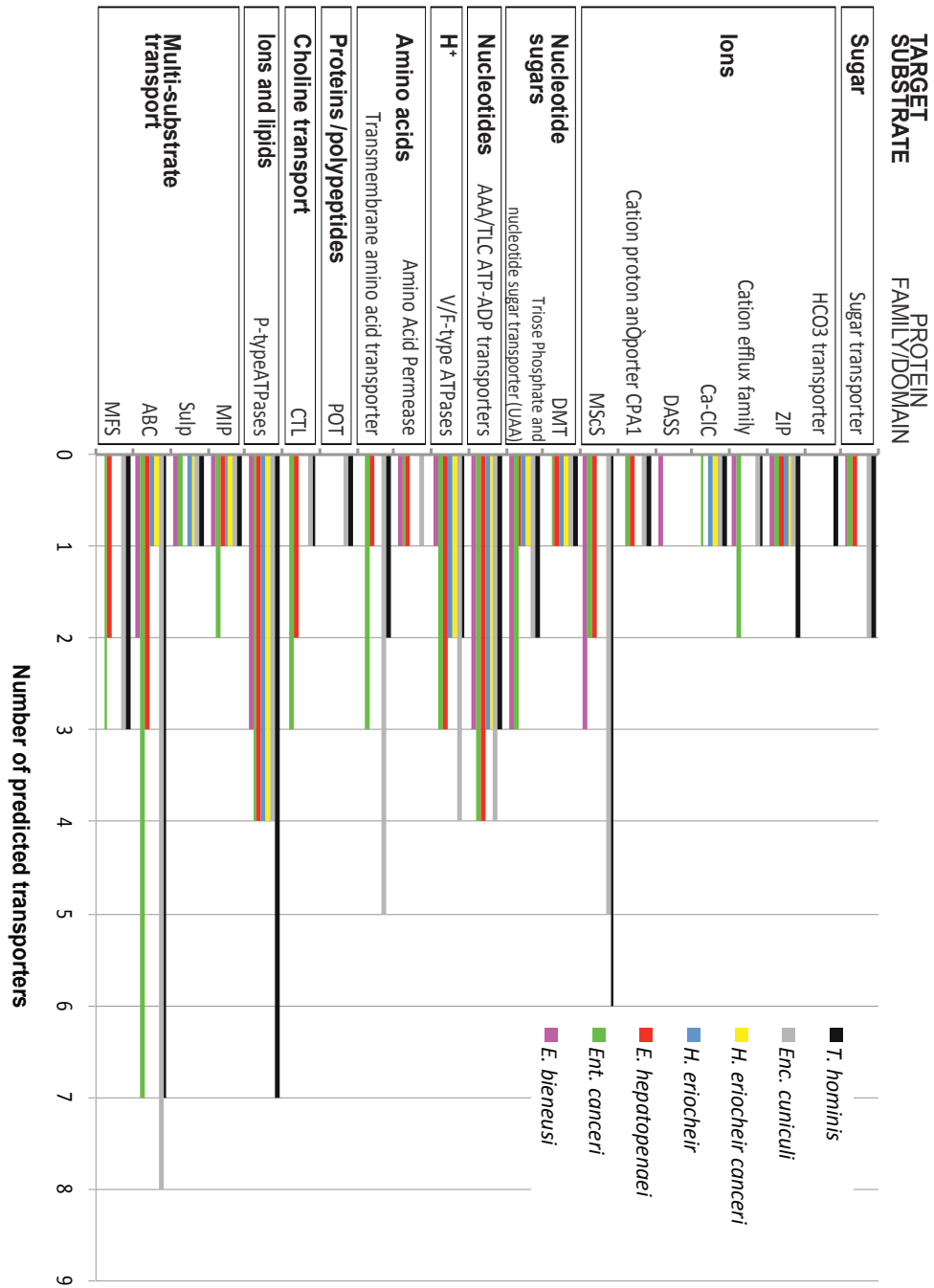


Figure 2.18: Number of predicted plasma membrane transporters belonging to various protein families identified for seven microsporidian genomes.

Table 2.19: Distribution of predicted plasma membrane transporters not assigned to a protein family

	<i>E. bieneusi</i>	<i>Ent. canceri</i>	<i>E. hepatopenaei</i>	<i>H. eriocheir</i>	<i>H. eriocheir canceri</i>	<i>Enc. cuniculi</i>	<i>T. hominis</i>
Paralogs unique to species	4	0	0	0	0	0	0
Orthologs but only found within the Enterocytozoonidae	0	1	2	1	1	0	0
Non-orthologous	10	0	1	0	0	0	1
Orthologs found in Enterocytozoonidae and in non-Enterocytozoonidae	128	105	85	92	82	111	107

2.4.13 Secreted proteins in the Microsporidia

An average of 149 secreted proteins with a standard deviation of 70 were predicted for the microsporidian phylum. Genomes of the *Encephalitozoon* clade had fewer predicted secreted proteins compared to the rest of the phylum with that of *Enc. romaleae* displaying only 61 secreted proteins. Conversely, *Edhazardia aedis*, the mosquito-infecting parasite displayed the highest number of predicted secreted proteins, 350. Within the Enterocytozoonidae family, the lowest and highest number of secreted proteins were found in the genomes of *H. eriocheir canceri* and *E. hepatopenaei* respectively (Figure 2.19) (Appendix 6). Genomes of closely related taxa such as *Hepatospora* spp., *Encephalitozoon* spp. and *Nematocida* spp. contained fewer unique and paralogous copies as compared to other members of the phylum. The majority of the predicted secreted proteins were hypothetical proteins. The few proteins annotated by MAKER or the BLAST2GO pipeline as non-hypotheticals have been listed in Table 2.20. Interestingly, no orthologous protein cluster was predicted to contain secreted proteins from all five Enterocytozoonidae species analysed in this study (Table 2.21). Thirteen orthologous protein clusters were identified to contain proteins likely to be secreted exclusively by more than one member of the Enterocytozoonidae (Table 2.21). Among secreted proteins predicted in this study, only one ortholog family was unique to all four crustacean infecting members of the Enterocytozoonidae (Table 2.21).

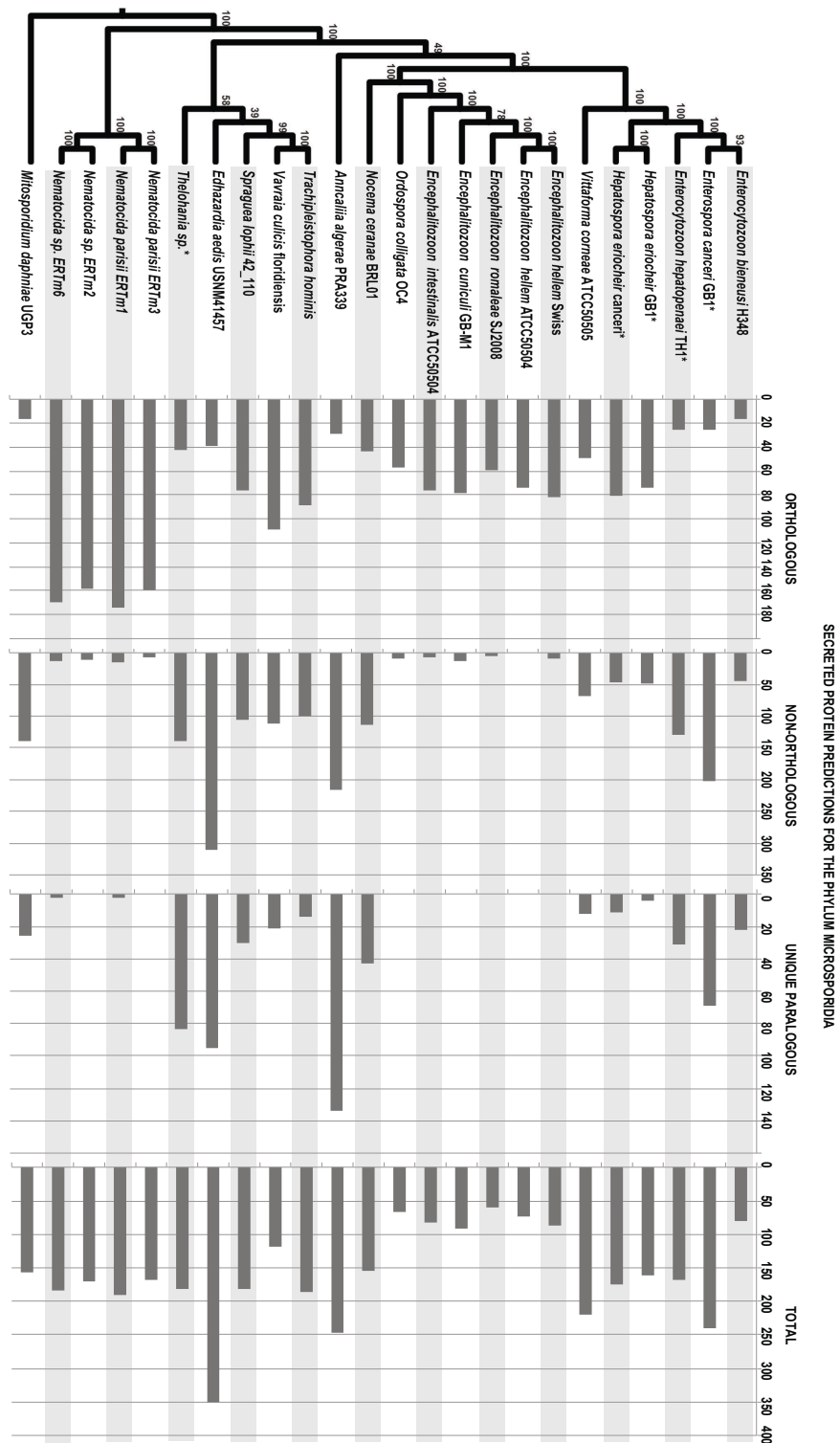


Figure 2.19: Comparative assessment of the number of orthologous, non-orthologous and unique-paralogous gene copies predicted to encode secreted proteins across the phylum Microsporidia. Taxa with * are newly sequenced taxa presented in this study. The phylogenetic positions in the cladogram are derived from maximum likelihood analysis performed on a concatenated alignment of 21 conserved proteins. Node values represent levels of support from 100 bootstrap replicates.

Table 2.20: Predicted secreted proteins annotated by MAKER and/or BLAST2GO

<i>Enterocytozoon hepatopenaei</i>	<i>Maker annotation</i>	<i>Blast2go top hit</i>
EHP00_151	hypothetical protein	zinc finger domain-containing
EHP00_393	hypothetical protein	WD40 repeat-containing
EHP00_372	Coatomer, beta subunit	WD40 domain-containing
EHP00_1424	hypothetical protein	transcriptional activator GLI3 isoform X1
EHP00_2597	hypothetical protein	TOLL1A
EHP00_1408	Cell cycle serine/threonine-protein kinase CDC5/MSD2	serine threonine kinase
EHP00_750	hypothetical protein	Serine hydroxymethyltransferase
EHP00_1010	hypothetical protein	lipase
EHP00_1718	hypothetical protein	lipase
EHP00_2316	hypothetical protein	lipase
EHP00_2325	rhoGAP protein	leucine-rich repeats and calponin homology (CH) domain containing 4
EHP00_532	pxylp1 (phosphoxylose phosphatase 1)	leucine-rich repeats and calponin homology (CH) domain containing 4
EHP00_761	Pigl (phosphatidylinositol glycan anchor biosynthesis)	leucine-rich repeats and calponin homology (CH) domain containing 4
EHP00_2619	hypothetical protein	leucine-rich repeats and calponin homology (CH) domain containing 4
EHP00_2611	hypothetical protein	leucine-rich repeats and calponin homology (CH) domain containing 4
EHP00_2613	hypothetical protein	leucine-rich repeats and calponin homology (CH) domain containing 4
EHP00_1967	hypothetical protein	leucine-rich repeats and calponin homology (CH) domain containing 4
EHP00_1979	hypothetical protein	leucine-rich repeats and calponin homology (CH) domain containing 4
EHP00_751	hypothetical protein	GNAT family acetyltransferase
EHP00_1942	hypothetical protein	endoplasmic reticulum membrane
EHP00_1308	Endochitinase	endochitinase
EHP00_771	hypothetical protein	cAMP-dependent kinase catalytic subunit
EHP00_944	SWP7 (Spore wall protein)	ABC-type multidrug transport ATPase and permease component
EHP00_2365	two-component system sensor histidine kinase	hypothetical protein
EHP00_1094	MIMI_L316 (glucosamine 6-phosphate N-acetyltransferase)	hypothetical protein
EHP00_1097	mel-32 (serine hydroxymethyltransferase)	hypothetical protein
EHP00_1920	LRR1 (leucine rich repeat protein)	hypothetical protein
EHP00_2050	Leucine-rich repeat protein	hypothetical protein
EHP00_2091	Leucine-rich repeat protein	hypothetical protein
EHP00_231	ATG15 (putative lipase)	hypothetical protein
<i>Enterospora cancri</i>	<i>Maker annotation</i>	<i>Blast2go top hit</i>
ECANGB1_1765	URK (uridine kinase)	uridine kinase
ECANGB1_915	AURKA (serine/threonine-protein kinase)	serine threonine- kinase
ECANGB1_2688	hypothetical protein	serine threonine kinase
ECANGB1_1052	hypothetical protein	serine threonine kinase
ECANGB1_1949	hypothetical protein	serine threonine kinase
ECANGB1_101	hypothetical protein	serine threonine kinase
ECANGB1_2129	hypothetical protein	nudix hydrolase
ECANGB1_177	hypothetical protein	lipase
ECANGB1_2678	hypothetical protein	endoplasmic reticulum membrane-associated oxidoreductin
ECANGB1_1632	CHIA (acidic endochitinase)	endochitinase

ECANGB1_2794	hypothetical protein	DUF1242 domain-containing
ECANGB1_1141	DNAJ (Hsp40)	DNA-binding protein
ECANGB1_447	hypothetical protein	disulfide isomerase
ECANGB1_2149	STK25 (serine threonine kinase)	cyclin-cyclin-dependent kinase 6
ECANGB1_2004	OMH1 (O-glycoside alpha-1,2-mannosyltransferase)	alpha-1,2 mannosyl-transferase
ECANGB1_2701	hypothetical protein	Alkaline protease 1
ECANGB1_2681	hypothetical protein	ABC-type multidrug transport ATPase and permease component
ECANGB1_1606	PTH2 (peptidyl-tRNA hydrolase)	hypothetical protein
ECANGB1_651	MKCA (protein kinase)	hypothetical protein
ECANGB1_239	FORF (formin)	hypothetical protein
ECANGB1_2529	transcription initiation factor TFIID subunit TAF9	hypothetical protein
<i>Hepatospora eriocheir</i>	<i>Maker annotation</i>	<i>Blast2go top hit</i>
HERIO_28	GGT (gamma-glutamyltranspeptidase)	gamma-glutamyltranspeptidase
HERIO_927	UDP-N-acetylhexosamine pyrophosphorylase	hypothetical protein
HERIO_1578	peptidyl-prolyl cis-trans isomerase	hypothetical protein
HERIO_2358	N-acetylglucosaminyl-phosphatidylinositol de-n-acetylase	hypothetical protein
HERIO_627	CTD small phosphatase	hypothetical protein
HERIO_272	chitinase 4-like	hypothetical protein
<i>Hepatospora eriocheir canceri</i>	<i>Maker annotation</i>	<i>Blast2go top hit</i>
A0H76_2998	hypothetical protein	dolichyl-phosphate-mannose- mannosyltransferase 1
A0H76_2994	hypothetical protein	DNA mismatch repair
A0H76_1756	CYP1 (peptidyl-prolyl cis-trans isomerase)	cyclophilin type peptidyl-prolyl cis-trans isomerase
A0H76_3025	hypothetical protein	acylase
A0H76_1595	UAP1 (UDP-N-acetylglucosamine pyrophosphorylase)	hypothetical protein
A0H76_2609	SWP7 (Spore wall protein)	hypothetical protein
A0H76_657	PSR2 (protein serine phosphatase)	hypothetical protein
A0H76_1065	PIGL (phosphatidylinositol glycan anchor biosynthesis)	hypothetical protein

Table 2.21: Functional annotation of orthoclusters of secreted proteins unique to 2 or more members of the Enterocytozoonidae

ORTHOCLUSTER ID	# genes in original cluster	# taxa in original cluster	<i>E. bienewisi</i>	<i>Ent. canceri</i>	<i>E. hepatopenaei</i>	<i>H. eriocheir</i>	<i>H. eriocheir canceri</i>	BLST2GO/BLAST Predicted function
1041	21	21	EBI_26397		EHP00_1094			MIMI_L316 (glucosamine 6-phosphate N-acetyltransferase)
137	34	23	EBI_26652				A0H76_496	RNA 3'-terminal phosphate cyclase
3015	4	3		ECANGB1_2751 ECANGB1_2752	EHP00_2126			no function predicted
1138	20	17		ECANGB1_1202 ECANGB1_2689	EHP00_266	HERIO_627	A0H76_657	PSR2 (protein serine phosphatase)
110	39	24		ECANGB1_177	EHP00_208			lipase
2238	5	3		ECANGB1_2702	EHP00_832			no function predicted
3412	3	2		ECANGB1_986	EHP00_1825			no function predicted
3430	3	3		ECANGB1_2034	EHP00_1702			no function predicted
5580	2	2		ECANGB1_359	EHP00_2133			no function predicted
5589	2	2		ECANGB1_2698	EHP00_2124			no function predicted
5598	2	2		ECANGB1_2207	EHP00_1483			no function predicted
5605	2	2		ECANGB1_2060	EHP00_751			no function predicted
5653	2	2		ECANGB1_1051	EHP00_1118			no function predicted

2.5 Discussion

2.5.1 Variability in assembly quality is due to different heuristic approaches used by assembly programs to assess errors, inconsistency and ambiguity.

The SPADES and A5-MISEQ assemblers have been repeatedly reported to outperform VELVET and RAY in past studies (Magoc et al. 2013; Gurevich et al. 2013; Utturkar et al. 2014; Koren et al. 2014; Chen et al. 2014; Mirebrahim et al. 2015; Jeong et al. 2015; Horn et al. 2016). It was therefore not surprising to observe similar results from both assemblers for the *H. eriocheir canceri* genome (Figure 2.1). Here, the considerably lower N50 values recorded for VELVET and RAY assemblies were due to short redundant contigs they produced that were fragments of larger contigs within the assembly. With regards to VELVET this problem may have been due to the absence of an inbuilt error correction Illumina read pre-processing step to correct erroneous base calls in reads (Zerbino & Birney 2008).

Although there is no evidence to suggest that the pre-processing error correction steps implemented in SPADES (BAYESHAMMER) and A5-MISEQ (TAGDUST/SGA) are directly responsible for their better assembly performance, there has been a past study where VELVET assemblies improved dramatically by the incorporation of extra pre-processing steps (Salzberg et al. 2012). In theory, the relatively high error rate characterized by Illumina sequencing data with respect to Sanger sequencing would mean de Bruijn graph-based assembly programs used to assemble Illumina data would benefit immensely from a read pre-processing step. This is because minimal abundance of base calling errors in reads will lead to the construction of less problematic de Bruijn graphs (i.e. graphs with minimal bubbles and tips) which will eventually lead to the construction of longer contigs (Zerbino & Birney 2008; Feldmeyer et al. 2011).

RAY does not perform initial pre-processing steps but uses a greedy algorithm on de Bruijn graphs. Thus, RAY builds contigs by using a de Bruijn graph to extend the sequence from regions of high coverage called seeds and terminates the extension if a set of kmers do not clearly indicate the direction of the extension. That is, unlike VELVET, SPADES and A5-MISEQ, RAY has very low tolerance for sequencing errors and terminates contig extension in regions of minimal error rate. This consequently leads to the assembly of only short, high

coverage contigs (Boisvert et al. 2010) which is reflected in the low N50 values recorded for the genome of *H. eriocheir canceri*. Another factor that may have contributed to the fragmentation of RAY's assembly is that RAY extracts k-mer information in splay tree structure which means that on a 64-bit computer (such as the one used for these assemblies), k-mer values can only go up to 32 (Boisvert et al. 2010). The inability of RAY to use longer k-mer values means that it is more likely to come across repetitive regions longer than the k-mer itself thereby creating an unresolvable ambiguity in the de Bruijn graph which terminates the extension of the contig (Boisvert et al. 2010).

An added feature in SPADES that may have contributed to its improved assemblies is its ability to automatically perform assembly iterations using different k-mer sizes and amalgamating the final individual assemblies into a final consensus assembly (Bankevich et al. 2012). As such the correct position of repetitive regions are recovered by longer k-mers thereby reducing the occurrence of contig extension termination as resulting from unresolvable ambiguities in the de Bruijn graph. However, longer k-mers have a relatively small coverage (as there is a reduced chance for the occurrence of another overlapping k-mer) and are more likely to introduce misassemblies in the final contigs (Zerbino & Birney 2008). To circumvent this problem where non-contiguous sequences are wrongly concatenated into long contigs, SPADES uses assembly iterations performed with smaller k-mer values to correct misassemblies (Bankevich et al. 2012). On top of the pre-processing step in the A5-MISEQ and SPADES pipelines, they also have a mismatch correction step where reads are remapped onto the assembled contigs to correct erroneous paths taken by the de Bruijn graph (Bankevich et al. 2012; Coil et al. 2015). This step is also used to concatenate contigs into scaffolds by harnessing the "insert size" information provided by paired-end Illumina reads. Considering the added error correction and contig extension steps provided by A5-MISEQ and SPADES which reflected in their improved performance, the selection of their assemblies as final assemblies for the microsporidian genomes sequenced in this project is justified (Bankevich et al. 2012;Coil et al. 2015).

2.5.2 Origin of contamination in Illumina data

DNA for whole genome sequencing (WGS) projects of currently published microsporidian genomes have so far been obtained from the spore stages

(Katinka et al. 2001; Fraser-Liggett 2005; Corradi et al. 2009; Corradi et al. 2010; Cuomo et al. 2012; Pombert et al. 2012; Heinz et al. 2012; Pan et al. 2013; Campbell et al. 2013; Pombert et al. 2013; Haag et al. 2014; Pombert et al. 2015). This is mainly because, for most species of this phylum, laboratory propagation is impossible and the spore is the only stage that can be isolated from environmental samples at levels high enough to pool down sufficient genomic DNA for sequencing. With respect to DNA extraction, the spore provides an excellent natural physical barrier to separate the desired microsporidian DNA from host's or other environment material. Thus, the spore's resilient nature makes it ideal for rigorous purification protocols. This is particularly important, as microsporidians are obligate intracellular parasites (Keeling & McFadden 1998; Keeling & Fast 2002; Heinz et al. 2012; Cuomo et al. 2012) (Section 1.4.1).

Apart from their intracellular habitat, members of the Enterocytozoonidae family sequenced in this study possessed small spore sizes and infected crustacean hepatopancreases, a niche well known for its rich microbial biodiversity (Bateman et al. 2011; Stentiford et al. 2012; Longshaw et al. 2012; Sweet & Bateman 2015). These factors presented an added difficulty for the extraction of pure genomic material for WGS. The spore sizes of *Ent. canceri*, *E. hepatopenaei* and *H. eriocheir* are 1.4 x 0.7 µm, 1.1 x 0.7 µm and 1.8 x 0.9 µm respectively (Stentiford et al. 2007; Stentiford et al. 2011; Tangprasittipap et al. 2013), which puts them in the size range of most bacterial species (Bakken & Olsen 1989; Kubitschek 1990). It was therefore not surprising to find high levels of bacterial contamination in the sequencing data for all three sequenced microsporidian species despite the rigorous spore purification protocols sequencing material were subjected to (Section 2.3.3-4). Bacterial contamination in microsporidian WGS projects is however common and has particularly been a problem in the assembled genome of *E. bienersi* (the only member of the Enterocytozoonidae family with a published genome) (Corradi et al. 2007; Akiyoshi et al. 2009; Cuomo et al. 2012; Heinz et al. 2012; Campbell et al. 2013; Pombert et al. 2013).

2.5.3 Low quality base pairs in raw reads of *Enterospora canceri* is as a result of poor library preparation

Illumina paired-end read library typically starts with the fragmentation of the double stranded DNA sample, size selection of desired fragments and ligation of the 5' ends of the fragments to adaptors composed of 6-12 bp unique

oligonucleotide sequences (Borgström et al. 2011). The adapters anchor the DNA fragment to the base of the flow cell and are used to initiate a series of bridge amplification cycles to generate monoclonal clusters. The amplification of double stranded fragments into monoclonal clusters is important for the production of high-intensity fluorescence during downstream sequencing-by-synthesis steps. Here, fluorescently labelled nucleotides and polymerases are used to simultaneously synthesise the monoclonal clusters for a desired length (read length), typically between 50-300 bp. During synthesis, an optical device records the incorporation of each nucleotide in real-time from each end of the double stranded DNA fragment resulting in paired-end reads. As such the initial raw read would be sequenced together with a 5' adapter DNA that needs to be removed prior to assembly.

Library preparation protocols are however not efficient at separating fragmented sample DNA into sizes, which occasionally leads to “read-through” errors. These errors occur in cases where DNA fragments in the library are shorter than the specified read length thereby causing sequencers to read through the adapter and inadvertently adjoining its sequence onto the 3' end of the reads (Caporaso et al. 2012). This seems to have been the case particularly for *Ent. canceri*'s RUN1 and 2 sequencing data as reads had long 3' stretches of poor quality DNA. Trimming of these contaminating adapter oligonucleotides led to the drastic reduction of the read length thereby negatively affecting the quantity of data presented to assembly programs.

Most assemblers are integrated with Illumina read trimming algorithms however, these programs rely on high scoring alignments between the Illumina reads and a dataset of known adapters for this purpose. Other programs use reverse complementarity between the 5' and 3' ends of the read to identify contaminating adapter sequences. Since the 3' ends of Illumina sequences have significantly higher error rates, identification of adapter sequences by these methods are not the most effective (Caporaso et al. 2012). Incorporation of untrimmed adapter sequences can significantly increase the time required for an assembler to run and also introduce misassemblies in to the final contigs (Bolger et al. 2014; Li et al. 2015). As such the FASTQC-informed-manual trimming approach implemented in this study (Section 2.3.7.1) was crucial in ensuring optimal performance of the assembly programs used. Selection of the assembly performed with reads from RUN3 as the final genome was because Illumina

reads from RUN3 had the best FASTQC read quality statistics. That is, whereas reads from RUN1 and 2 needed to be trimmed in order to improve their assembly quality, RUN3 reads did not. Furthermore, the assembly performed with reads from RUN3 displayed the highest mean coverage (328 X) compared to assemblies from RUN1 and 2 (16 and 134 X respectively).

2.5.4 Decontamination protocols depend on target and contaminating genome properties

2.5.4.1 *Hepatospora eriocheir canceri*: A case of low target and high contaminant genomic GC content

Genomes of obligate intracellular parasites are often associated with low GC content and microsporidian genomes are no exception (Rocha & Danchin 2002; Williams et al. 2008; Merhej et al. 2009; Akiyoshi et al. 2009; Cuomo et al. 2012; Campbell et al. 2013). In this study, this characteristic of microsporidian genomes was harnessed to eliminate likely contamination present in the sequencing data. GC content filtration of the Illumina reads was particularly effective in the optimization of the genomic assemblies of *H. eriocheir canceri*. The improvement of the assembly statistics, especially the increase in length of the assembly's largest contig after GC filtering suggests that contaminant reads have a negative impact on the performance of assembly programs (Table 2.2). Unsurprisingly, the assembly of the unfiltered reads required higher volumes of RAM space and took more processing time than that for filtered reads. As such applying this filtering step to the raw reads rather than at the contig level proved more efficient. The negative effect of contaminant reads and the effectiveness of read GC filtering in improving assemblies has also been highlighted in recent studies (Kumar et al. 2013; Zhou & Rokas 2014).

2.5.4.2 *Enterocytozoon hepatopenaei*: A case of low target and contaminant genomic GC content

For the assembly of *E. hepatopenaei*'s genome, inspection of the read coverage across the assembly and BLAST analysis revealed that this assembly was contaminated with bacterial DNA. Most of the contaminating contigs had strong BLAST hits to *Acinetobacter* bacterial species and a GC content of 35.2 %, a value that is similar to that of the published *Acinetobacter baumannii* genome, 39.2 % (Park et al. 2011). This relatively low GC content of the *Acinetobacter* sp.

contaminating genome explains its intractableness towards the low GC filtering cut-off values implemented in this study (Figure 2.4). The presence of *Acinetobacter* sp. in the sampled shrimp materials could be explained by the fact that *Acinetobacter* spp. share the same aquatic habitat with the shrimp host of *E. hepatopenaei* (Lee & Pfeifer 1977; Petersen et al. 2002; Agersø & Petersen 2007). Even though *Acinetobacter* spp. are naturally free-living saprophytes, some species are known to be opportunistic parasites and symptomatic shrimp for Slow Growth Syndrome are often plagued with high levels of bacteria (Flegel 2012). Since the shrimp sampled for the extraction of *E. hepatopenaei* spores were symptomatic for Slow Growth Syndrome, they may have harboured high levels of *Acinetobacter* infections. Also, these bacterial cells are approximately 1.3 x 1.0 µm in size, which puts them in the size-range of *E. hepatopenaei*, 1.1 x 0.7 µm (James et al. 1995; Tangprasittipap et al. 2013). Their small size may have enabled a few bacterial cells to be pooled together with *E. hepatopenaei* spores thereby explaining the presence of their genomic DNA in the initial assembly, 2.2 Mbp (Figure 2.5).

The Percoll density gradient filtration step employed to purify spores in this study would have however made it less likely for the contaminating bacterial cells to pool in high quantities with microsporidian cells despite their small cell size. As a result of this, there would have been less contaminating bacterial DNA present in the genomic DNA sample submitted for sequencing than microsporidian DNA. This difference in DNA abundance between the two organisms would have in turn reflected in the Illumina sequencing read coverage. Thus the most abundant organism in the genomic DNA sample submitted for sequencing would show high read coverage whereas the less abundant organism would show low read coverage. Unsurprisingly, almost all contigs identified as bacterial sequences had a very low coverage whereas microsporidian contigs had high coverage (Figure 2.5).

2.5.4.3 *Enterospora canceri*: a case of high GC content in target and contaminant genome

Microsporidian genomes are known to be GC poor with *Nosema* species presenting values as low as 18.78 % (Chen et al. 2013). Other lineages such as *Enc. cuniculi*, *Enc. intestinalis*, *Enc. Hellem* and *Vav. culicis* have relatively higher GC content (Figure 1.6) (Katinka et al. 2001; Corradi et al. 2010; Pombert et al.

2012; Haag et al. 2014; Desjardins et al. 2015). An initial assessment of *Ent. canceri*'s assembled contigs identified microsporidian genes with a GC content higher than 40 %. Unlike the genomes of *H. eriocheir canceri* and *E. hepatopenaei* where the total number of core microsporidian genes present in the assembled genome plateaued with increasing GC cut-offs, this value increased sharply with increasing GC-cutoff values and reached a maximum only in the unfiltered read assembly (Figure 2.7). This made GC content filtration unfavourable for the optimization of this assembly. Furthermore, unlike the case of *E. hepatopenaei* where contigs of contaminating organisms presented with low read coverage, the unimodal coverage distribution observed for the RUN1 assembly (Figure 2.6) suggested either: There was no difference in the read coverage between contaminant and target contigs or there was no contamination present in the assembly.

The second premise was however unlikely since assemblies performed with reads from all three sequencing RUNs possessed the same number of core microsporidian genes and yet RUN2 and 3 presented smaller assemblies. This implied that the contigs present in RUN1 but absent in RUN2 and 3 were contaminants. This was confirmed by BLAST results which demonstrated that majority of the unfiltered RUN1 contigs had non-microsporidian BLAST hits (Appendix 11). Since read coverage distribution was unable to distinguish between contaminating and microsporidian contigs, k-mer coverage, contig length and BLAST searches were used as parameters for this task. These proved to be invaluable parameters for filtering out likely contaminating contigs (Figure 2.8 and 2.10). Considering *Ent. canceri* is an intranuclear parasite, presence of host material such as rDNA in the initial assembly was not surprising (Figure 2.8). Due to the highly repetitive nature of crustacean rDNA, such contaminating contigs showed high k-mer coverage (Figure 2.8). The rigorous purification protocols performed on the microsporidian spore samples prior to DNA extraction and library preparation for sequencing (Section 2.3.3-4) meant that contaminating genomes that persisted in the initial assembly would be incomplete and therefore very fragmented. This was observed in Figure 2.8 where all contaminating contigs were below 15 kbp. The k-mer and read coverage improvement of the second assemblies, which were constructed from reads that mapped back onto the filtered contigs corroborates previous studies that showed that contaminating reads have a negative impact on the performance of assembly programs (Alkan

et al. 2010; Schmieder & Edwards 2011; Kumar et al. 2013). In order to recuperate microsporidian sequences present in only RUN1 or RUN2, filtered reads from both RUNs were combined to perform a joint assembly (Table 2.15). Although the resulting assembly was less fragmented, it was at least ~1 Mbp shorter than the assembly performed with reads from RUN2.

Due to the poor library preparation for RUN1 & 2 and the high levels of contaminating reads from RUN 1, new genomic DNA was extracted from spores and submitted for sequencing, RUN3. Following initial BLAST-based purification steps, this sequencing run produced better quality reads which translated in a better assembly (Table 2.16). The assembly resulting from the amalgamation of reads from all three RUNs had worse statistics compared to those performed on reads from individual RUNs (Table 2.17). Perhaps this is because the combination of reads from all three RUNs also meant combining errors inherent in each RUN. The final assembly submitted to GENBANK was that from RUN3.

2.5.5 Phylogenomics of the Enterocytozoonidae

Ribosomal DNA-based phylogenetic assessment of the microsporidians currently assigned to the Enterocytozoonidae family place *Ent. canceri* as the closest relative of the human parasite *E. bienewisi* (Figure 1.4). Whereas this observation is corroborated by previous studies (Freeman et al. 2013; Palenzuela et al. 2014), contradictions also exist in the literature where *E. hepatopenaei* have been placed as the closest relative to the human infecting parasite (Tourtip et al. 2009; Stentiford, Feist, et al. 2013), all three parasite branch as a distinct clade (Stentiford et al. 2011) or *E. bienewisi* branches as a separate clade (Freeman & Sommerville 2009). Such contradictions have had a profound effect on the systematics of the Enterocytozoonidae. Examples include the case of the copepod parasite, *Desmozoon lepeophtherii* that was assigned a second taxonomic name by another research group (Freeman & Sommerville 2009; Nylund et al. 2010; Freeman & Sommerville 2011) (Section 1.8.1.1). Such examples demonstrate the current confusion in the systematics of this group of microsporidians.

In an effort to shed more light on the phylogenetic relationships within the Enterocytozoonidae, the new genomic data created in this study in addition to genomic data for microsporidian species currently available on the public database, MICROSPORIDIADB (Aurrecochea et al. 2011) were harnessed to

construct a 21-protein concatenated phylogenetic tree (Figure 2.15). *Thelohania* sp. sequenced in this study was also added to the analysis. The topology of the multi-protein tree was identical for both maximum likelihood and Bayesian Inference methods employed in this study (Figure 2.15). Similar tree topologies have also been published in previous studies (James et al. 2013; Nakjang et al. 2013). Taking this, and the high statistical support recorded for the Enterocytozoonidae clade into account, there is strong evidence to suggest that *Ent. canceri* is the closest relative to *E. bieneusi* and the *Hepatospora* spp. are indeed the earliest branching species in this clade. Although the topology of this tree is bound to change in the future as a result of increased availability of WGS data for presently unsequenced member species of the Enterocytozoonidae, the phylogenetic relationship between *E. hepatopenaei*, *Ent. canceri* and *E. bieneusi* is unlikely to change. This data strongly support the assignment of the genus name, *Enterospora* to the shrimp parasite, *Enterocytozoon hepatopenaei*.

2.5.6 Enterocytozoonidae and transposable elements

In canonical fungi, transposable elements (TEs) can have a considerable contribution to the overall genome size (Daboussi & Capy 2003) and this has also been demonstrated in the Microsporidia (Williams et al. 2008; Corradi et al. 2009). That is microsporidian lineages with bigger genomes such as *Anncaliia algerae* (Peyretailade et al. 2012), *Hamiltosporidium tvaerminnensis* (Corradi et al. 2009), *Edhazardia aedis* (Williams et al. 2008) and *Trachipleistophora hominis* (Heinz et al. 2012) have been documented to possess high levels of transposable elements as compared to their counterparts with smaller genomes (Parisot et al. 2014). Two different programs, harnessing different databases (DFAM and REPBASE) and search algorithms (Hidden Markov Models and RMBLAST) were employed to investigate the presence of TEs in genomes sequenced in this study and that of *E. bieneusi*. Interestingly both programs showed the presence of TEs in all Enterocytozoonidae genomes however, REPEATMASKER, the BLAST-based program, identified a smaller contribution for TEs in the investigated genomes when compared to DFAMSCANNER. For *E. bieneusi*, 2.5 % of the genome length was predicted to be represented by TEs by DFAMSCANNER whereas only 0.05 % was predicted by REPEATMASKER.

The poor performance of BLAST-based programs in identifying TEs was also reflected in a recent study that reported no presence of TEs in the genome of *E.*

bieneusi (Parisot et al. 2014). In this study, the authors tried to identify TEs in *E. bieneusi* by doing a TBLASTX search with homologues from *A. algerae*.

The majority of the TEs predicted for the genomes of microsporidia sequenced in this study were situated in either intergenic regions, hypothetical genes or chaperone coding genes. There is a strong likelihood that many of the identified TEs are in fact false positives. However, the prediction of TE by both programmes and the presence of transposase and reverse transcriptase-coding genes in the genomes of *E. bieneusi* (EBI_22056 and EBI_24469), *Hepatospora* spp. (A0H76_1644, HERIO_2566) and *Ent. canceri* (ECANGB1_1040, ECANGB1_1564 and ECANGB1_429) is suggestive of a real occurrence of TEs in the Enterocytozoonidae. Interestingly, all the reverse transcriptases identified in the genome of *Ent. canceri* have strong BLAST hits to avian sequences, suggesting the horizontal transfer or sequence contamination from an avian host. Considering some avian species prey on the crab hosts of *Ent. canceri* (Sibly & McCleery 1983), a horizontal transfer of TEs from an avian species into *Ent. canceri* is a possibility. Secondly, the low prevalence (1/200 crabs) of *Ent. canceri* in crabs (Stentiford & Bateman 2007) is suggestive of the presence of a reservoir host which may well be an avian species. In the future, screening of seabird pellets with *Ent. canceri* specific primers designed from genomic data presented here could be used to test this hypothesis.

2.5.6.1.1 Gypsy and Tc1 retrotransposons: Ubiquitous transposable elements in the Microsporidia

As with TEs found in sequenced microsporidian genomes to date, the majority of the TEs observed in the genomes of *Ent. canceri*, *E. hepatopenaei* and *Hepatospora* spp. belonged to the gypsy and Tc1 superfamily (Cornman et al. 2009; Peyretailade et al. 2012; Parisot et al. 2014). A TE superfamily commonly represented in other microsporidian genomes but was absent in the genomes of *Ent. canceri*, *E. hepatopenaei* and *Hepatospora* spp. was piggyback (Parisot et al. 2014). This however is not surprising as there is evidence suggesting the recent horizontal acquisition of members of this TE superfamily from invertebrate hosts in the Microsporidia (Heinz et al. 2012). TEs identified in the Enterocytozoonidae genomes sequenced in this study constituted only a small part of their overall genome size (0.3-3 %).

2.5.7 Absence of complementary tRNAs for some codons in microsporidian ORFeomes

A conspicuous difference between the codon-usage-bias versus GC distribution plots of ORFs for taxa sequenced in this study and *E. bieneusi* is that ORFs belonging to *E. bieneusi* did not aggregate in a defined region of the graph (Figure 2.14). Instead, *E. bieneusi* ORFs displayed a more scattered codon-usage-bias versus GC distribution, which is strongly suggestive of assembly contamination - an observation also highlighted by other investigators (Heinz et al. 2012). The presently studied Enterocytozoonidae have similar GC contents in their first and second codon positions but the GC distribution at their third, synonymous codon position varied (Figure 2.14). For example, *Ent. canceri*'s ORFs displayed a strong preference for G or C nucleotides at their third codon positions (Figure 2.15). That is, in instances where multiple codons coded for the same amino acid, *Ent. canceri* consistently displayed a higher preference for those with a G or C nucleotide at the third codon position (Figure 2.15). Results from this study are consistent with previous findings that show there is a positive correlation between low genomic GC composition and low GC composition at third codon positions (Hershberg & Petrov 2008; Chen et al. 2004; Cornman et al. 2009).

It has been previously observed that abundance of certain tRNAs can influence synonymous codon usage such that a genome adapts to utilize the most abundant tRNAs (Ikemura 1985; Akashi & Eyre-Walker 1998; Duret 2002; Lavner & Kotlar 2005). No such pattern was observed in the genomes sequenced in this study and in *E. bieneusi*. That is in instances where high levels of a tRNA with a particular anticodon is encoded by a genome, a higher representation of its complementary codon was not observed in the organism's ORFeome (Figure 2.12) (Figure 2.15).

Another observation in taxa sequenced in this study is that not all codons in their ORFeomes had complementary tRNAs encoded by their respective genomes (Figure 2.12). This observation is however ubiquitous across the different domains of life and has been demonstrated to be as a result of "wobble base-pairing" (Chan & Lowe 2009). This is described as the scenario where tRNAs present within a cell pair with synonymous codons that have a single non-complementary nucleotide at their third codon position. The genomes of *E. bieneusi* and the *Hepatospora* spp. coded for 41/61 tRNAs whereas those of *Ent. canceri* and *E. hepatopenaei* coded for 38/61 tRNAs, even fewer than what is

encoded by the *E. coli* genome, 39/61 (Chan & Lowe 2009) (Figure 2.12). All sequenced genomes contained fewer tRNAs than the complete genome of *Enc. cuniculi* (retrieved from MICROSPORIDIADB) which contained 44/61 tRNAs. For a phylum renowned for their reduced genomes as the Microsporidia, a condensed tRNA repertoire capable of forming wobble base pairing is not surprising and may even be responsible for increasing translational speed as recently described in yeast (Gardin et al. 2014).

2.5.8 The GGGTAAAA motif: A putative transcription binding site of the Enterocytozoonidae

Previous studies have hinted to the likely importance of a GGG or CCC followed by two nucleotides and an AAAATTTT sequence as transcription binding sites in the Microsporidia (Cornman et al. 2009; Peyretailade et al. 2009). In the Enterocytozoonidae genomes analysed in this study, the trailing “TTTT” was however less conserved and even absent in the genomes of *E. bieneusi* and *E. hepatopenaei* (Figure 2.11). This observation that this group of microsporidia have condensed some of their transcription binding sites is indeed interesting as it adds to the list of genomic items that the Microsporidia have jettisoned/compacted in order to condense their genomes. No transcription binding sites predicted in this study were exclusive to all members of the Enterocytozoonidae although the CTTCTT and AAAATA sequences were exclusive to subsets within the family (Figure 2.11). Apart from GGGTAAAA sequence mentioned above, there were three more sequences identified as common transcription factor binding sites within the taxa analysed in this study, which included the TATA box domain (Figure 2.11). This domain has also been identified for the genome of *Nosema* sp. (Cornman et al. 2009). In summary, these data suggest that the GGGTAAAA domain may be an important transcription binding signal in the Enterocytozoonidae and that individual taxa within this microsporidian family may be evolving taxa-specific transcriptional signals. The latter perhaps reflects to an extent the different hosts and subcellular environments these taxa inhabit. These putative transcription factor-binding sequences could be an important resource for the validation of gene predictions for new members of the Enterocytozoonidae in future annotation projects, as has been shown by Peyretailade et al., 2009.

2.5.9 Extreme reduction in metabolic capacity within the Enterocytozoonidae

Whilst individual microsporidian lineages have lost enzymes of different metabolic pathways independently, the Enterocytozoonidae have undergone the most dramatic metabolic reduction described to date. In addition to the loss of most glycolytic enzymes, members of this family have also lost genes involved in the pentose phosphate pathway, fatty acid and trehalose metabolism (Figure 2.17).

2.5.10 Partial conservation of deoxyribonucleotide metabolism within the Microsporidia

In the cell, deoxyribonucleotides can be synthesized either via a salvage or *de novo* pathway. The salvage pathway initially attaches purine and pyrimidine rings onto activated ribose sugars via phosphoribosyltransferases to form ribonucleotide mono-phosphates (NMPs). Following this, NMPs are phosphorylated by monophosphate kinases (URA6) to ribonucleotide diphosphates (NDPs). NDPs are then converted into deoxyribonucleotides diphosphates (dNDPs) by ribonucleotide reductase (RNR2). This is followed by the phosphorylation of the dNDPs to fully charged deoxyribonucleotide triphosphates (dNTPs) by a nucleotide diphosphate kinase (YNK1). This enzyme is also responsible for the conversion of NDPs to ribonucleotide triphosphates (NTPs), which are the building blocks for RNA with ATP in particular having the extra role of being an energy currency within the cell (Berg et al. 2006) (Figure 2.20).

De novo deoxyribonucleotide metabolism on the other hand employs different pathways for purine and pyrimidine synthesis. Pyrimidine synthesis involves the assembly of pyrimidine rings from bicarbonate, aspartate and glutamine precursors followed by the coupling of these rings onto an activated ribose sugar to form NMPs. Purine synthesis involves the assembly of purine rings from precursors directly onto a ribose sugars to form NMPs. These partially charged ribonucleotides are converted into fully charged dNTPs via a phosphorylation reaction followed by a reduction reaction similar to that described above for the salvage pathway (Berg et al. 2006) (Figure 2.20).

All 23 microsporidians genomes analysed in this study possessed only those genes involved in the latter phase of deoxyribonucleotide metabolism (See thick

arrows in Figure 2.20) and not genes responsible for the early stages of this pathway. These microsporidians are therefore unlikely to undertake either salvage or *de novo* deoxyribonucleotide metabolism. Furthermore, the enzyme ribose phosphate diphosphokinase (see PRS1 in Figure 2.20), which activates ribose sugars in order to prepare them for purine and pyrimidine ring recruitment is absent in all analysed microsporidia. This enzyme was only found in the genome of *M. daphniae*, a close relative of the microsporidia with a publicly available draft genome (Haag et al. 2014). Interestingly, these observations have never been made in any published comparative genomic analyses performed in the past.

All microsporidian genomes analysed in this study encoded at least three transporters belonging to the TLC/AAA transporter family (Figure 2.20) (Appendix 10). Members of this group of transporters from *T. hominis* have been shown to be effective in importing purines: ATP, GTP, ADP and GDP and not pyrimidines: UTP, CTP, UDP and UTP when expressed in *E. coli* (Heinz et al. 2014). The conservation of both purine and pyrimidine salvage pathways in all analysed microsporidia is however suggestive of the presence of a yet unidentified transporter responsible for importing pyrimidines within the microsporidia (Figure 2.20). In the future, functional characterization of transporter proteins not assigned to any protein family identified in this study could be useful in identifying this enigmatic pyrimidine transporter.

Taken together, the data suggest that the Microsporidia have lost the intrinsic capacity of nucleotide production but have acquired specific transporters capable of nucleotide import from host. They have also maintained the enzyme repertoire necessary for converting partially charged NMPs, NDPs, dNMPs and dNDPs to fully charged NTPs and dNTPs (Figure 2.20). Although this nucleotide-recycling pathway is capable of ATP generation, it seems unlikely for ATP produced in this manner to be used to drive metabolic processes as it is energetically expensive: It takes 2 ATP molecules to convert AMP to ATP (Berg et al. 2006; Mathews 2015). The conservation of the latter phase of deoxyribonucleotide metabolism across the microsporidia is likely for the purposes of converting imported NDPs and NTPs into dNTPs thereby making these DNA precursors readily available for DNA duplication during the proliferative merogonial life stage.

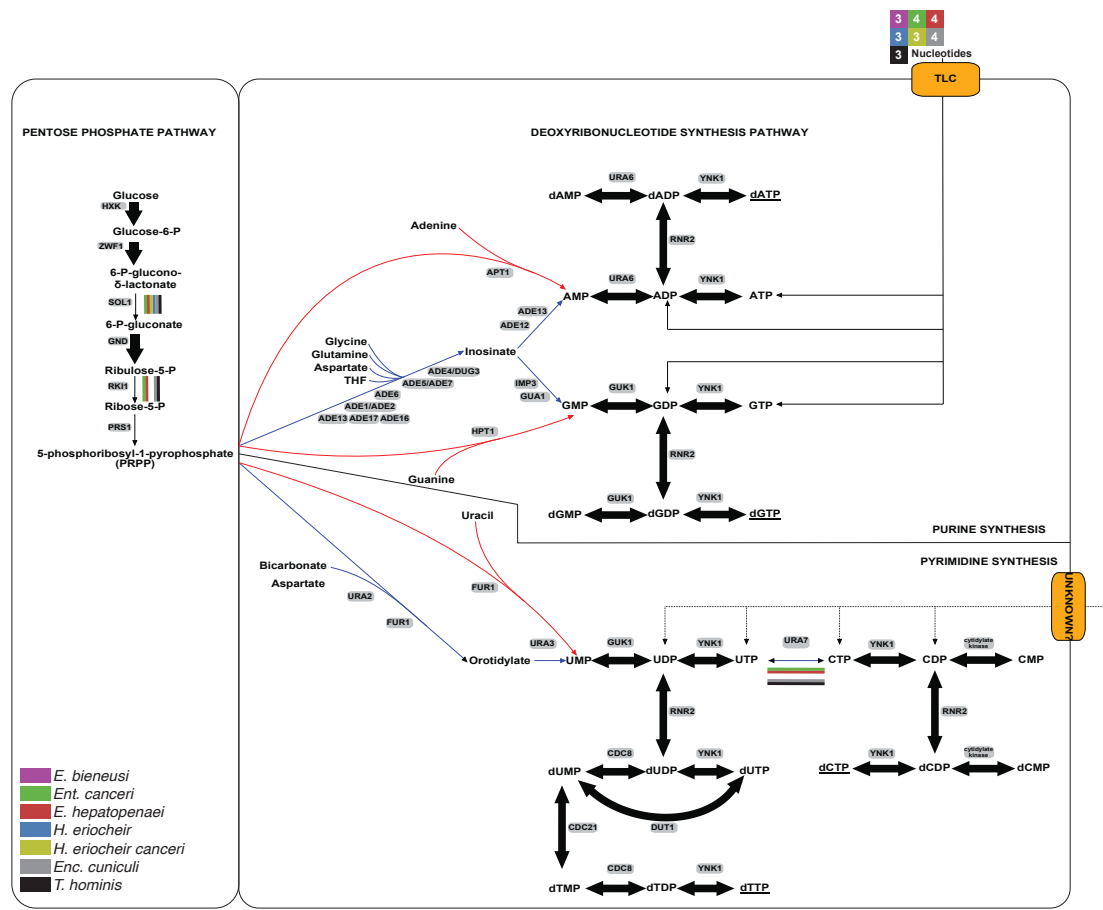


Figure 2.20: Comparing the enzyme repertoire of the pentose phosphate and deoxyribonucleotide synthesis pathway between members and non-members of the Enterocytozoonidae family (figure adapted from Nakjang et al. 2013) Homologs of *Saccharomyces cerevisiae* metabolic genes (represented on grey background) from the 23 microsporidians used in this analysis were retrieved by orthology clustering. For the purposes of simplicity only 2 non-Enterocytozoonidae species (*Encephalitozoon cuniculi* and *Trachipleistophora hominis*) were represented in this figure. Thick black arrows represent those pathways found in all Enterocytozoonidae and the two non-Enterocytozoonidae genomes selected. Where there is differential loss of the pathway, presence of a gene in each species is represented by a different coloured line. Predicted number of the TLC-type plasma membrane transporter family are colour coded for the seven species represented in this figure. “Unknown” in the figure represents yet uncharacterised plasma membrane transporters likely to be responsible for the import of pyrimidines in the Microsporidia. dNTPs are underlined.

2.5.11 Assessment of plasma membrane transporters to uncover a signature of intranuclear parasitism

The Microsporidia, as with most obligate intracellular parasites, have relinquished many core metabolic processes and rely on their host for the acquisition of key metabolites through specialised plasma membrane transporters (Katinka et al. 2001; Tsoulos et al. 2008; Williams et al. 2008; Heinz et al. 2012; Nakjang et al. 2013). The heavy dependence of the Microsporidia on these transporters makes these proteins attractive therapeutic targets. Considering the members of the Enterocytozoonidae are well known for causing disease in economically important fisheries and humans, identification of such transporters in this clade is an important step towards development of therapeutic strategies. In this study,

sequences of putative plasma membrane transporters have been identified from new genomic data for *Ent. canceri*, *E. hepatopenaei*, *H. eriocheir*, and *H. eriocheir canceri* (Appendix 10). This will be an invaluable database for directing research in the development of pharmacological targets for these parasites.

Among the transporter families represented in microsporidian genomes analysed here, multi-substrate, nucleotide, nucleotide sugars, and ion and lipid transporters had the most representatives. The retention of multi-substrate ABC transporters across all analysed taxa highlights the importance of transporters with promiscuous substrate specificity in organisms with reduced transporter repertoire such as the Microsporidia (Dean et al. 2014). The scant metabolic repertoire characterised by member taxa of the Enterocytozoonidae presented in this study (Figure 2.17), which is more severe than what has been observed in published microsporidian genomes (Heinz et al. 2012) suggests an even greater reliance of members of this clade on their plasma membrane transporters. In addition, I hypothesised that the niche intranuclear environment parasitized by the *Ent. canceri* will reflect in its plasma membrane transporter repertoire. To investigate this, the complement of plasma membrane transporters across the Enterocytozoonidae were compared to determine whether the intranuclear *Ent. canceri* genome encoded different transporters to sister taxa inhabiting the cytoplasm. Interestingly, no transporters with known functions were predicted to be exclusive to the *Ent. canceri* genome. However, the genome of this parasite encoded more transporters with a predicted function than any other Enterocytozoonidae members (Appendix 10). This study therefore suggests that *Ent. canceri* may have expanded its transporter repertoire in order to support its intranuclear lifestyle. In future studies, functional characterization of transporters not assigned to any protein family identified in this study (Appendix 10) will be key in understanding how *Ent. canceri* survives in the nucleus of its host.

2.5.12 The predicted secretomes of *Hepatospora* spp., *Enterocytozoon*

bieneusi*, *Enterospora canceri* and *Enterocytozoon hepatopenaei

Considering the number of proteins encoded by microsporidian genomes analysed in this study, secreted proteins make up 2-8 % of the total number of proteins encoded by microsporidian genomes (Appendix 6). This value is comparable to that reported for other intracellular parasites such as *Plasmodium*

falci-parum (5 %) (Hiller et al. 2004), *Leishmania donovani* (2 %) (Silverman et al. 2008) and *Trypanosoma brucei* (4 %) (Geiger et al. 2010). As expected, closely related microsporidian taxa, infecting similar hosts, such as *Nematocida* and *Encephalitozoon* species had minimal or no non-orthologous (unique) secreted protein groups (Figure 2.19). Interestingly, this was not observed for members of the Enterocytozoonidae (Figure 2.19) (Appendix 6), raising doubts on whether members of this family such as *E. bienewisi* and *Ent. canceri* are as closely related as portrayed by phylogenetic analysis (Figure 2.16). It is however important to note that members of this clade, have a broad host range (Chalifoux et al. 1998; Haro et al. 2006; Stentiford & Bateman 2007; Tourtip et al. 2009; Stentiford et al. 2011; Zhao et al. 2015). This means that secreted proteins from a human infecting parasite such as *E. bienewisi* would be under completely different evolutionary pressures compared to those of crab infecting parasites such as *Ent. canceri*. Furthermore, the different host subcellular niches these parasites inhabit may play a key role in the evolution of secreted proteins. These factors will in turn lead to the expansion, loss or diversification of certain secreted protein families to make the parasite more fit for its environment. Consequently, high numbers of unique secreted protein families within the Enterocytozoonidae at levels not observed in other closely related species is perhaps not entirely unexpected. The high numbers of unique secreted proteins in the genomes of *H. eriocheir* and *H. eriocheir canceri* (78 and 98 respectively) was however unexpected as these parasites are members of the same species in multi-gene phylogenetic analyses (Chapter 5, also in Bateman et al. 2016 see Appendix 12). A close inspection of predicted ORFs for these genomes demonstrated that although a few proteins were truly absent from both assemblies, they both contained a similar protein repertoire but sequencing errors and fragmented contigs had altered a subset of predicted ORFs in both genomes. These alterations made it difficult for the orthology clustering program to group these proteins into orthologous families. An interesting observation is that no orthologous family of secreted proteins was observed to be ubiquitous across all 23 microsporidian genomes or all 5 members of the Enterocytozoonidae family analysed in this study (Appendix 6). Due to their involvement in manipulating the host environment, secreted proteins need to evolve in response to evolutionary pressures from the host such as immune responses (Chisholm et al. 2006). Since these evolutionary pressures are host specific and the Microsporidia infects a plethora of hosts, the absence of

orthologous secreted protein families across the Microsporidian phylum or across the Enterocytozoonidae is also not surprising. Even though no protein cluster contained genes from all five representative taxa of the Enterocytozoonidae, one protein cluster contained proteins from four representative taxa of this family. This orthocluster included proteins annotated as serine\threonine phosphatase (Table 2.21). The function of this protein family is currently not fully understood but bacterial parasites such as *Porphyromonas gingivalis* are known to secrete these proteins into their host environment to repress host immunity (Takeuchi et al. 2013) and to invade the host cell (Tribble et al. 2006). Amongst the other 11 orthoclusters unique to the Enterocytozoonidae only 3 were assigned a function by BLAST2GO/BLAST analysis (Table 2.21). One of these clusters contained proteins annotated as lipases. This protein family was particularly abundant in *E. hepatopenaei* (Table 2.20). Lipases have been identified as effectors in other microbes and are implicated in virulence, transmission, life cycle development, modulation of host lipids and host immune responses (DeAngelis et al. 2007; Sikora et al. 2011; Nascimento et al. 2016). Members of the other two orthoclusters were annotated as glucosamine 6-phosphate N-acetyltransferase and RNA 3'-terminal phosphate cyclase. Both of these proteins have not been documented as effector proteins in other parasites but have been implicated in chitin biosynthesis (Kato et al. 2006) and DNA and RNA ligation (Chakravarty & Shuman 2011) respectively. Without any functional data, it is difficult to establish the role of these presumptive secreted proteins in parasite-host interaction.

2.6 Conclusion

Apart from *E. bieneusi*, all other known taxa within the Enterocytozoonidae are pathogens of aquatic animals (Stentiford, Feist, et al. 2013). Here, the genomes of four of these taxa known to infect aquatic crustaceans have been sequenced. Considering *Hepatospora* spp. are the earliest branching members of currently known Enterocytozoonidae, the public availability of their genomic data will provide an important resource for rooting phylogenomic trees aimed at this clade. Furthermore, this study provides the first genome sequence for the major yield-limiting shrimp pathogen *E. hepatopenaei*. Its public availability will undoubtedly underpin the development of new tools to diagnose, monitor and potentially mitigate the negative impacts of this pathogen. The genome of *Ent. canceri* represents the first from an obligate intranuclear eukaryotic pathogen and in particular. The study has provided an invaluable resource for the investigation of host-pathogen interaction, systematics, genetic reduction and pathogen evolution in arguably the most economically and evolutionarily interesting family within the Microsporidian phylum (Stentiford et al. 2016). Some of the analyses performed in this chapter have been submitted for publication to the Journal of Environmental Microbiology (See Wiredu-Boakye *et al.*, 2016 in Appendix 13).

Chapter 3 The phylum Microsporidia and loss of glycolytic enzymes

3.1 Introduction

3.1.1 Universality of glycolysis and plasticity within this pathway

Glycolysis, also known as the Embden-Meyerhof-Parnas pathway is a ubiquitous pathway across the tree of life often relied upon by cells for the conservation of energy and generation of reducing potential in the form of ATP and NADH respectively. (Fothergill-Gilmore & Michels 1993; Berg et al. 2006) (Figure 3.1). In eukaryotes, genes involved in this pathway are thought to have been transferred from the primordial endosymbiont, α -proteobacteria (which later became the mitochondria) into the nuclear genome of its methanogenic archaean host (Canback et al. 2002; Wu et al. 2004). Despite its ancient origin and integral role in both prokaryotic and eukaryotic cells, the glycolytic pathway is characterized by extreme plasticity in many lineages. For example, the last glycolytic enzyme pyruvate kinase, is one of the two enzymes responsible for the release of ATP in this pathway. However, in the bacterial parasite *Treponema pallidum* this enzyme is absent and the parasite relies on a yet uncharacterized enzyme for the conversion of phosphoenolpyruvate (PEP in Figure 3.1) to pyruvate (Das et al. 2000). *Helicobacter pylori* bypasses its setback of not possessing a pyruvate kinase gene by using an alternative glycolytic pathway, Entner-Doudoroff pathway (ED), amino acid and oligopeptide fermentation for energy conservation (Schilling et al. 2002). *Mycobacterium bovis*, bypasses the problem of its inactive pyruvate kinase by depending on fatty acid catabolism for its energy needs and not depending on glycolysis at all (Keating et al. 2005) (Figure 3.1).

Glycolytic plasticity in eukaryotes is hallmarked by horizontal procurement of various glycolytic genes from prokaryotes. A typical example is the glucokinase gene of intracellular parasites, *Giardia intestinalis*, *Trichomonas* spp., *Trypanosoma* spp. and *Spironucleus barkhanus* that appear to originate from a prokaryotic ancestor (Henze et al. 2001; Wu et al. 2001; Cáceres et al. 2007). Free-living protists such as *Naegleria gruberi* also possess a eubacterian-like glucokinase. *Trimastix pyriformis*, a free-living nanoflagellate has horizontally acquired almost half of its glycolytic genes set from an ancestral eubacteria

(Stechmann et al. 2006). Similar to the above-mentioned scenario in the *T. pallidum* bacterium, eukaryotic lineages such as *N. gruberi*, *Entamoeba histolytica*, *G. lamblia* and *T. vaginalis* catalyze the final step of glycolysis with a pyruvate kinase isoenzyme called pyruvate-phosphate dikinase. This enzyme transforms the unidirectional irreversible conversion of phosphoenolpyruvate to pyruvate into a reversible reaction thereby enabling the conversion of excess pyruvate into free glucose (Reeves 1968; Hrdý et al. 1993; Opperdoes et al. 2011). These examples demonstrate that despite the universality of glycolysis, plasticity in this pathway is inherent in both parasitic and non-parasitic prokaryotes and eukaryotes. Furthermore these examples demonstrate that horizontal acquisition of glycolytic genes has occurred on several occasions in both free living and parasitic eukaryotes. Microsporidia as a phylum seems to have benefited from several horizontal gene transfer events, however none of the horizontally acquired genes identified in the Microsporidia to date are glycolytic genes: ATP/ADP antiporter, H⁺/nucleoside symporter, piggyback DNA transposons, catalase, manganese superoxide dismutase, class II photolyase, GTP cyclohydrolase I, folic acid synthase, phosphoribosyltransferase and purine nucleotide phosphorylase (Richards et al. 2003; Slamovits & Keeling 2004; Tsaousis et al. 2008; Lee et al. 2009; Xiang et al. 2010; Pombert et al. 2012; Cuomo et al. 2012; Heinz et al. 2012; Pan et al. 2013; Watson et al. 2015).

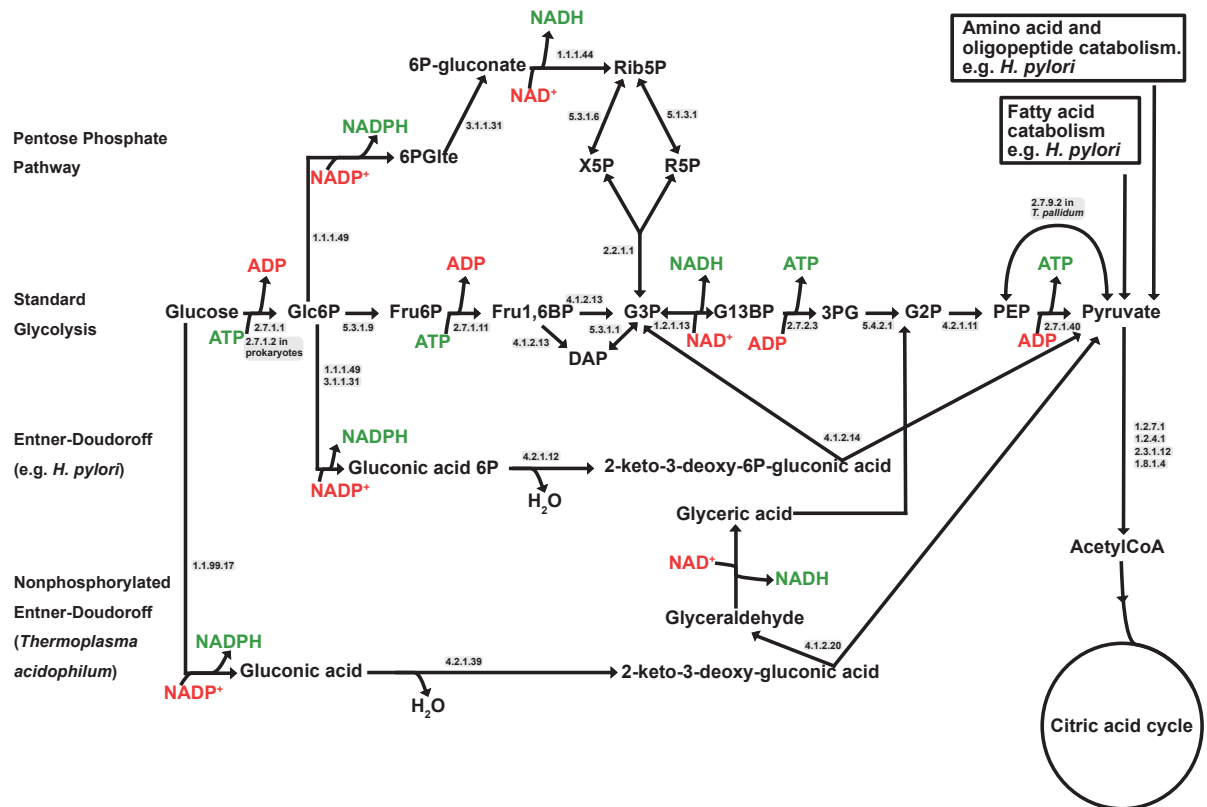


Figure 3.1: Schematic representation of the glycolytic pathway (centre) and alternative routes: Entner-Doudoroff and Nonphosphorylated Entner-Doudoroff pathway found in various species (in brackets). A pathway for the breakdown of glucose involving the pentose phosphate pathway has also been shown. Arrows symbolize the substrate flux of the enzymatic reactions. The release or consumption of energy or reducing potential is indicated in green or red respectively. Enzymes that catalyze various reactions are represented by their EC numbers: Hexokinase (EC 2.7.1.1), glucokinase (EC 2.7.1.2), phosphoglycerate mutase (EC 5.4.2.1), phosphoenolpyruvate synthase (EC 2.7.9.2); phosphoglycerate kinase (EC 2.7.2.3), glucose 6-phosphate 1-dehydrogenase (EC 1.1.1.49), 6-phosphogluconolactonase (EC 3.1.1.31), phosphogluconate dehydrogenase (decarboxylating) (EC 1.1.1.44), 6-phosphogluconate dehydrogenase decarboxylating (EC 1.1.1.44), D-ribulose 5-phosphate 3-epimerase (EC 5.1.3.1), ribose-5-phosphate isomerase (EC 5.3.1.6), transketolase (EC 2.2.1.1), glucose dehydrogenase (EC 1.1.99.17), gluconate dehydratase (EC 4.2.1.39), 2-dehydro-3-deoxyglucuronate aldolase (EC 4.1.2.20), pyruvate kinase (EC 2.7.1.40). Abbreviations: PEP, phosphoenolpyruvate; DAP, dihydroxyacetone phosphate; Fru6P, fructose 6-phosphate; Fru1,6BP, fructose 1,6-bisphosphate; Glc6P, glucose 6-phosphate; gluconic acid 6P, 6-phosphogluconic acid; G3P, 3-phosphoglycerate; G13BP, 1,3-bisphosphoglycerate; G23BP, 2,3-bisphosphoglycerate; G2P, 2-phosphoglycerate; Rib5P, ribulose 5-phosphate; R5P, ribose 5-phosphate; X5P, xylulose 5-phosphate; 6PGIte, 6-phosphoglucono- δ -lactonate. Image adapted from Dandekar et al., 1999.

3.1.2 How the Microsporidia acquire ATP

3.1.2.1 Glycolysis

The breakdown of glucose in the Microsporidia is thought to follow the standard glycolytic pathway also known as the Embden-Meyerhof pathway (Williams et al. 2014) (Figure 3.1). Here, a single glucose molecule is broken down into pyruvate in a series of reactions often catalyzed by 10 enzymes (Figure 3.1). This process requires the use of 2 ATP molecules but releases four ATP molecules thereby

producing a net of 2 ATP molecules per glucose molecule. In eukaryotes, the end product of glycolysis, pyruvate is further catabolized to release 26 ATP molecules via the citric acid cycle and the oxidative phosphorylation pathway (Berg et al. 2006). Microsporidians are however an exception as all genomes sequenced till date are devoid of genes involved in the citric acid cycle and the oxidative phosphorylation pathway (Katinka et al. 2001; Fraser-Liggett 2005; Corradi et al. 2009; Corradi et al. 2010; Cuomo et al. 2012; Pombert et al. 2012; Heinz et al. 2012; Pan et al. 2013; Campbell et al. 2013; Pombert et al. 2013; Haag et al. 2014; Pombert et al. 2015). Immunoblotting studies of glycolytic genes of the different microsporidian life stages demonstrated that key glycolytic genes such as glycerol-3-phosphate dehydrogenase and phosphoglycerate kinase are highly expressed in the cytosol of spores and not in the proliferative merogonial stages (Dolgikh et al. 2011; Heinz et al. 2012). In the light of these findings and the likelihood of spore germination to be ATP dependent (Weidner & Byrd 1982), some authors have suggested that glycolysis is mainly used for ATP generation during the spore stage, and this ATP is used to fuel spore germination (Williams et al. 2014). Considering the likely importance of glycolysis in spore germination, it was indeed surprising to find that the genome of the human parasite, *Eneterocytozoon bieneusi* did not encode most glycolytic enzymes (Akiyoshi et al. 2009; Keeling et al. 2010).

3.1.2.2 Horizontally acquired ATP/ADP translocases

The current consensus on energy acquisition in the intracellular, proliferative merogonial stage is that they primarily rely on ATP acquisition from the host cytosol via horizontally transferred ATP/ADP translocases situated on their plasma membrane (Richards et al. 2003; Tsaousis et al. 2008). There is evidence to suggest that horizontal acquisition of ATP/ADP translocases occurred in the common ancestor of the Microsporidia and Cryptomycota and that the single horizontally acquired translocase has undergone several species-specific duplications and deletions during the evolutionary history of extant microsporidian lineages (Heinz et al. 2014). A copy of this gene within *Encephalitozoon cuniculi* seems to have evolved specialized functions as it has been shown to localize specifically to the parasite's mitosome (Tsaousis et al. 2008). The significance of this is perhaps to provide ATP for the microsporidian mitosome during iron-sulphur cluster biosynthesis (Goldberg et al. 2008) but the absence of mitosomal

targeting homologs in other microsporidian species suggests this specialization may be unique to *Enc. cuniculi* (Heinz et al. 2014). Apart from the mitochondrial targeting ATP/ADP translocase, *Enc. cuniculi* possess 3 more homologs of the protein that localizes to its plasma membrane (Tsaousis et al. 2008).

3.1.3 The microsporidian hexokinase

Hexokinase is the first enzyme of the glycolytic pathway and plays the crucial role of catalyzing the phosphorylation of cytosolic glucose molecules into glucose-6-phosphate (Glc6P in Figure 3.1) by the transfer of a phosphoryl group from ATP onto the glucose molecule (Berg et al. 2006). This prevents glucose to be transported out of the cell via glucose transporters situated on cell membranes. The phosphorylation of glucose into Glc6P therefore commits glucose molecules to progress down a number of pathways such as glycolysis, pentose phosphate pathway, trehalose metabolism and chitin synthesis (Berg et al. 2006; Williams et al. 2014)(Figure 3.2). Apart from its role in glycolysis, hexokinase is involved in the synthesis of mannose-containing glycoconjugates which are needed for cell wall and membrane synthesis (Davis & Freeze 2001; Patterson et al. 2003; T. Chen et al. 2014). Thus, as its name suggests, hexokinase does not only catalyze the phosphorylation of glucose but also other hexose sugars such as mannose and fructose (Schnarrenberger 1990). Hexokinases that solely phosphorylate certain hexose sugars, often referred to as glucokinase or fructokinase have been documented in various lineages including prokaryotes (Albig & Entian 1988; Doehlert et al. 1988; Schnarrenberger 1990; Bork et al. 1993; Hansen et al. 2003). Hexokinase has also been documented to regulate gene expression in various model organisms including yeast and *Arabidopsis* (Rodríguez et al. 2001; Ahuatzzi et al. 2004; Cho et al. 2006). This multifaceted role played by hexokinase in cellular metabolism is aided by its occurrence in some genomes as multiple copies often referred to as isoforms (Katzen & Schimke 1965; Reeves et al. 1967; Entian et al. 1984; Entian & Fröhlich 1984; Mayordomo & Sanz 2001; Wolf et al. 2011). Equally, the genomes of microsporidians encode for multiple copies of hexokinase (Nakjang et al. 2013). This is however not observed across the entire phylum and phylogenetic analysis show that hexokinase has undergone lineage-specific expansions in the Microsporidia (Nakjang et al. 2013). There is currently no information available on whether the isoforms encoded by microsporidian genomes are involved in

intrinsic specialized functions as observed in yeast (Entian et al. 1984; Entian & Fröhlich 1984; Albig & Entian 1988). Some studies have however demonstrated that a number of microsporidian hexokinases may be specialized to exploit the host environment. That is some microsporidian hexokinases encode secretion signals and therefore may be secreted into the host to boost glycolytic production of ATP within the host which is subsequently pilfered by the microsporidian meront via specialized translocases (Tsaousis et al. 2008; Cuomo et al. 2012; Senderskiy et al. 2014).

In this chapter, the discovery made in chapter 2 of an a-glycolytic lifestyle for some members of the Enterocytozoonidae (Section 2.4.11) will be further investigated: A comparative phylogenomic approach will be employed on data generated in chapter 2 to confirm this finding and to examine the role of remnant glycolytic genes (hexokinases) in energy conservation and metabolic process in a-glycolytic microsporidian species. I hypothesise that glycolysis is indeed lost in *Ent. canceri* and that an a-glycolytic lifestyle may infact be a common trait among the Enterocytozoonidae family. I further predict that remnant glycolytic genes in a-glycolytic genomes may be nonfunctional.

3.2 Main aims of study

- To understand when and how glycolysis was lost in the microsporidian phylum.
- To understand the function of remnant glycolytic genes in members of the Enterocytozoonidae family

3.3 Methods

3.3.1 Bioinformatics

3.3.1.1 Hexokinase phylogeny

Identification of microsporidian hexokinases was performed by parsing the microsporidian orthology cluster file created in Section 3.3.3. The orthologous cluster containing the *Enc. cuniculi* protein annotated as hexokinase (ECU11_1540) on the MicrosporidiaDB public webserver (Aurrecoechea et al. 2011) was selected. The *Ent. canceri* hexokinase protein was retrieved by manually parsing through the BLASTP output as it clustered with the PTPA protein orthology cluster. The hexokinase orthologs for *Saccharomyces cerevisiae* (NP_116711.3) (NP_011261.1) (NP_009890.1), *Rozella allomycis* (EPZ31577.1), *Rattus rattus* (NP_036866.1) (NP_036867.1) (NP_001257778.1) and *Homo sapiens* (NP_000179.2) (NP_000180.2) (NP_000153.1) were extracted from the publicly available NCBI database (Tatusova et al. 2014) by querying the identified hexokinase gene of *Mitosporidium daphniae* against their respective databases. A total of 50 hexokinases were aligned with the command line version of MUSCLE (v3.8.31) (Edgar 2004). The most rigorous masking option on TRIMAL (v1.2rev59) (Capella-Gutierrez et al. 2009) was employed to mask the alignment by using the “-strictplus” option. The best substitution model for the masked protein was predicted with the PROTTEST (v3.4.2) command line tool (Abascal et al. 2005). The predicted GTR substitution model together with a GAMMA rate heterogeneity module was used on command line RaxML (Stamatakis 2014) to construct a maximum likelihood tree. A second phylogenetic analysis was performed with *Ent. canceri*'s hexokinase without its PTPA domain: The coordinates of the PTPA domain was identified by performing a search against the NCBI Conserved Domain database (Marchler-Bauer et al. 2015) and the PTPA domain was manually removed from *Ent. canceri*'s hexokinase protein sequence. The uncoupled hexokinase protein sequence was used for phylogenetic analysis following the above-described protocol. The removed PTPA domain was also used for separate phylogenetic analysis as described in Section 3.3.1.2 below.

3.3.1.2 PTPA phylogeny

Identification of microsporidian PTPA proteins was performed by parsing the microsporidian orthology cluster file created in Section 3.3.3. The orthologous cluster containing the *Ent. canceri*'s chimeric PTPA-hexokinase protein was selected. All positions within the alignment containing gaps and missing data were masked using the command line TRIMAL tool with the “-strictplus” option (Capella-Gutierrez et al. 2009). The evolutionary history was inferred by using the maximum likelihood and a discrete Gamma distribution was used to model evolutionary rate differences among sites. The command line version of RaxML (Stamatakis 2014) was used to infer the evolutionary history.

3.3.1.3 Mapping hexokinase active sites

Individual and stretches of amino acids of *S. cerevisiae*'s hexokinase 1 that were empirically identified by Kuser et al., 2008 as important for the normal functioning of the enzyme were manually identified from the hexokinase alignment described above. Pairwise alignments were performed between the identified active sites of each hexokinase used in this analysis and that of *S. cerevisiae* hexokinase 1 with the online BLAST2 tool (Mount 2007). Percentage sequence identity for each active site was recorded in to a CSV file and converted into a colour gradient in R (Ihaka & Gentleman 2012). That is, higher sequence identities were represented by stronger shades whereas weak sequence identities were represented by a lighter shades.

3.3.1.4 Genome-wide analysis to identify chimeric proteins within the *Microsporidia*

The predicted open reading frames ORFs for each of the assembled genomes and those of publicly available microsporidians were submitted to the online chimeric gene prediction tool, GENE DEFUSER (Salim et al. 2011).

3.3.1.5 Assessment of gene order

This was performed by doing a local BLASTP (Camacho et al. 2009) (e-value cutoff 0.001) search of all predicted microsporidian ORFs against their own database and passing the resulting BLAST file to the command line version of ORTHOMCL (Li et al. 2003) which group proteins into orthologous families. Contigs containing orthologous families were visualized using a genomic comparative tool, ACT (Carver et al. 2008).

3.3.2 Cloning: Organisms, strains and plasmids

3.3.2.1 Bacterial strains

The three bacterial strains used in this study are described in table 2.1 below.

Table 2.1 Strain names of *Escherichia coli* used, molecular features, sources and purpose of use.

Strain name	Molecular features	Source	Experiment
CcdB Survival 2	F- <i>mcrA</i> Δ (<i>mrr-hsdRMS-mcrBC</i>) Φ 80 <i>lacZ</i> Δ M15 Δ <i>lacX74 recA1 ara</i> Δ 139 Δ (<i>ara-leu</i>)7697 <i>galU galK rpsL</i> (StrR) <i>endA1 nupG fhuA::IS2</i>	Dr. Hsueh-Lui, University of Exeter	Propagating ccdB Gateway vectors: pDONR221 and pDEST17
TOP10	F- <i>mcrA</i> Δ (<i>mrr-hsdRMS-mcrBC</i>) Φ 80 <i>lacZ</i> Δ M15 Δ <i>lacX74 recA1 ara</i> Δ 139 Δ (<i>araleu</i>)7697 <i>galU galK rpsL</i> (StrR) <i>endA1 nupG</i>	Invitrogen, UK Ltd.	Propagating Gateway entry and expression clones: pENT221-microsporidianHexokinase, pEXP17-microsporidianHexokinase
Rosetta2(DE3)pLYS	F- <i>ompT hsdSB</i> (rB- mB-) <i>gal dcm</i> (DE3) pLysSRARE (CamR)	Millipore, UK Ltd.	Recombinant hexokinase gene expression

3.3.2.2 Yeast strains

The three yeast strains used in this study are described in table 2.1 below.

Table 2.2 Strain names of *Saccharomyces cerevisiae* used, molecular features, sources and purpose of use.

Strain name	Molecular features	Source	Experiment
W303-1A	MATa { <i>leu2-3,112 trp1-1 can1-100 ura3-1 ade2-1 his3-11,15</i> }	ThermoFisher Scientific, UK	Control for yeast complementation assay
YSH7.4-3C Triple hexokinase KO strain of W303-1A	MATa { <i>leu2-3,112 trp1-1 can1-100 ura3-1 ade2-1 his3-11,15</i> } <i>hk1</i> Δ <i>hxk2</i> Δ <i>glk1</i> Δ	(De Winde et al. 1996)	For yeast complementation assay
BY4741	MATa <i>his3</i> Δ 1 <i>leu2</i> Δ 0 <i>met15</i> Δ 0 <i>ura3</i> Δ 0	(Brachmann et al. 1998)	For heterologous expression of recombinant microsporidian hexokinases

3.3.3 Primers

For a comprehensive list of primers used in this study, refer to Appendix 3

3.3.4 Plasmids

For a list of plasmids used in this study and their respective detailed maps please refer to Appendix 2.

3.3.5 Media and solutions

3.3.5.1 Bacterial media solutions

The following media and solutions were used in this study.

Luria-Bertani (LB) broth, 1000 ml	<ul style="list-style-type: none">• 10 g Bacto-tryptone• 5 g yeast extract• 10 g NaCl• H₂O up to 1000 ml
LB-agar, pH 7.2, 1000 ml	<ul style="list-style-type: none">• 10 g Bacto-tryptone• 5 g yeast extract• 10 g NaCl• 10 g Agar• H₂O up to 1000 ml
SOC, 1000 ml	<ul style="list-style-type: none">• 20 g Bacto-tryptone• 5 g yeast extract• 2 ml of 5 M NaCl• 2.5 ml of 1 M KCL• 10 ml of 1 M MgCl₂• 10 ml of 1 M MgSO₄• 20 ml of 1 M glucose• H₂O up to 1000 ml
Kanamycin (50 mg/ml)	0.5 g of Kanamycin disulfate salt was dissolved in 10 ml H ₂ O and filter-sterilized through a 0.2 µm filter. Final concentration of 50 µg/ml was used.
Ampicillin (100 mg/ml)	1 g of sodium ampicillin was dissolved in 10 ml H ₂ O and filter-sterilized through a 0.2 µm filter. Final concentration of 100 µg/ml was used

3.3.5.2 Yeast media

Yeast Extract Peptone Dextrose media (YEPD)	<ul style="list-style-type: none">• 20 g Bacto-peptone• 10 g yeast extract• H₂O up to 1000 ml
YEPD-agar	<ul style="list-style-type: none">• 24 g Bacto-agar

Yeast Minimal/Synthetic Defined (SD) media	<ul style="list-style-type: none"> • 10 g yeast extract • 10 g Bacto peptone • H₂O up to 1000 ml • 20 g Glucose (for W303-1A)/ Galactose (for YSH7.4-3C) • 0.72 g Yeast nitrogen base without amino acids • 6.99 g complete supplement - URA dropout • H₂O up to 1000 ml
Yeast Minimal/Synthetic Defined (SD) agar	<ul style="list-style-type: none"> • 20 g Glucose (for W303-1A)/ Galactose (for YSH7.4-3C) • 24 g Bacto-agar • 0.72 g Yeast nitrogen base without amino acids • 6.99 g complete supplement - URA dropout <p>H₂O up to 1000 ml</p>

3.3.5.3 DNA electrophoresis

Tris/Acetic acid/EDTA (TAE) electrophoresis buffer 1000 ml	<ul style="list-style-type: none"> • 4.84 Tris base • 1.14 ml Glacial acetic acid • 0.29 g EDTA
1 % Agarose gel	<ul style="list-style-type: none"> • 5 g of Agarose • Top up to 500 ml with 1 X TAE • Microwaved at full power till Agarose is completely dissolved

3.3.5.4 Solutions for protein work

Running buffer	<ul style="list-style-type: none"> • 3.04 g Tris • 14.4 g Glycine • Topped up with H₂O to 1000 ml
Transfer buffer	<ul style="list-style-type: none"> • 3.04 g Tris

10 X PBS	<ul style="list-style-type: none"> • 14.4 g Glycine • 200 ml Methanol • Topped up with H₂O to 1000 ml • 4.56 g NaH₂PO₄ • 23 g Na₂HPO₄ • 87.66 g NaCl • Topped up with H₂O to 1000 ml
1 X PBS Tween-20 (0.01 %)	<ul style="list-style-type: none"> • 50 ml 10 X PBS • 50 µl Tween-20 • Topped up with H₂O to 500 ml
1 X PBS + 0.01% Tween-20 + 5 % milk	<ul style="list-style-type: none"> • 2.5 g milk • Topped up to 50 ml with 1 X PBS 0.01 % Tween
1 X PBS 0.01% + Tween-20 + 1 % milk	<ul style="list-style-type: none"> • 0.5 g milk • Topped up to 50 ml with 1 X PBS 0.01 % Tween
2XBinding/washing buffer	<ul style="list-style-type: none"> • 100 mM Sodium-phosphate, pH 8.0 • 600 mM NaCl • 0.02 % Tween-20 • Diluted to 1X for final use
His-elution buffer	<ul style="list-style-type: none"> • 300 mM imidazole • 50 mM Sodium-phosphate pH 8.0 • 300 mM NaCl • Tween-20

3.3.6 Molecular techniques

3.3.6.1 Addition of *attB* sites onto amplified microsporidian hexokinase

Primers GWE_ca-p1F/R, GWE_ca1F/R, GWp1F/R and GWH_e1F/R were used to amplify the hexokinase gene from their respective microsporidian genomes and add recombination sites from the bacteriophage lambda (*attB* sites) to the flanking ends of the gene (Appendix 3). In order to promote high fidelity sequence

amplification, Phusion High-Fidelity DNA Polymerase (New England BioLabs, UK) was used according to manufacture's instructions. PCR products were subsequently verified by gel electrophoresis.

3.3.6.2 Resolution of DNA fragments by gel electrophoresis

The success of each Gateway PCR amplification step was evaluated by running the PCR mixture on a 1 % agarose gel and checking for a DNA band of an expected size. Here, 1 µl of the PCR reaction mix was diluted in 5 µl of H₂O and 1 µl of loading dye (New England BioLabs, UK) prior to being loaded on the agarose gel. GoTaq Green Master Mix (Promega, UK) was only used for colony PCRs to identify positive transformants. 5 µl of the final PCR reaction was directly loaded on a 1 % agarose gel. Samples were run for 30 minutes at 80 V. DNA visualization was performed in a G:Box gel imager (Syngene, UK).

3.3.6.3 Purification of PCR products from agarose gels

The DNA gel was carefully placed on an open UV transilluminator (Wealtec, UK). The illumined DNA bands were carefully excised with a sharp sterile razor and placed in 1.5 ml Eppendorf tubes. Purification of PCR product from the excised agarose gels was performed by using the GenElute Gel Extraction Kit (Sigma, UK).

3.3.6.4 BP clonase reaction

Shuttling of the amplification products into the pDONR221 entry vector (Appendix 2) was performed following the Life Technologies BP clonase manual.

3.3.6.5 LR clonase reaction

Shuttling of the recombinant microsporidian hexokinase from the entry vector into a destination vector (Appendix 2) was performed following the Life Technologies LR clonase manual.

3.3.6.6 Chemical *Escherichia coli* transformation

50 µl of competent *Escherichia coli* cells were transferred into precooled Eppendorf tubes. 1 ng of plasmid was added to the aliquot of competent cells and mixed by gentle stirring. This mixture was left on ice for 30 minutes followed by a 30 second heat-shock in a 42 °C water bath. The heat-shocked cells were stood on ice for 2 more minutes before being diluted with 200 ml of SOC. The diluted competent cells were incubated for at 37 °C at 180 rpm for 1 hour to encourage

expression of the antibiotic resistance gene. The transformed cells were subsequently plated on LB agar plates supplemented with antibiotic for 16 hours and positive transformants were detected by PCR using the appropriate primers.

3.3.6.7 Isolation of plasmid DNA from *Escherichia coli* cells

Positive transformants were grown overnight in 50 ml of selective LB media at 37 °C at 180 rpm. Overnight incubation was harvested and plasmids extracted according to the Miniprep plasmid extraction kit protocol (Promega, UK).

3.3.6.8 Cloning poly-HIS tagged microsporidian hexokinases into PYES2 *Saccharomyces cerevisiae* vectors

Hexokinase gene sequences of *Enc. cuniculi* were used to design primers for a two-step PCR reaction to amplify the gene from genomic DNA, append a poly-his tag to the C-terminal of the translated protein and restriction enzyme sites on both ends of the sequence (Sacl and Xbal to the 5' and 3' ends of the gene respectively). Amplified PCR products were cloned using a TOPO TA cloning system by following manufacturer's instructions (Invitrogen, UK). Ligated vectors were transformed into TOP10 competent *E. coli* (Invitrogen, UK) using protocols in Section 3.3.6.6. Transformed colonies with PCR inserts were selected and grown overnight in LB/ampicillin medium, 100 µg/ml ampicillin. Overnight *E. coli* cultures were harvested for plasmid extraction and purification using the Qiagen miniprep plasmid extraction kit (Qiagen, UK), following manufacturer's protocols. Purified plasmids were double digested with NdeI and SacII restriction enzymes (New England BioLabs, UK) and the restriction fragment was separated from the plasmid by gel electrophoresis on a 1% agarose gel. The restriction fragment was subsequently excised and cleaned with Qiagen's gel extraction kit (Qiagen, UK) following manufacturer's manual.

Sequencing results for plasmid inserts were checked against their respective hexokinase gene sequences using the MUSCLE webserver (Edgar 2004) and aligned sequences were then viewed using SEAVIEW (Gouy et al. 2010). Upon ascertaining the amplified PCR fragments were the correct sequence, they were cloned into PYES2 yeast expression vectors (Invitrogen, UK) using the Sacl and Xbal restriction enzyme sites present on the vector. Standard sticky end ligation protocols provided with the T4 DNA ligase enzyme mix were followed (New England BioLabs, UK). The cloned plasmids were transformed into BY4741 strains, following protocols in Section 3.3.6.10 and plated on SD plates containing

appropriate selection and incubated at 30 °C for 5 days. Positive transformants were checked by PCR with T7 and CYC1 primers.

3.3.6.9 Cloning microsporidian hexokinases into pAG426GPD-EGFP

***Saccharomyces cerevisiae* vectors**

Primers GWE_ca-p1F/R, GWE_ca1F/R were used to amplify *Ent. canceri*'s hexokinase from its genome and adjunct recombination sites from the bacteriophage lambda (*attB* sites) to the flanking ends of the gene as explained in Section 3.3.6.1. Resolution of DNA fragments by gel electrophoresis, purification of PCR products and BP clonase reaction were performed as explained in Section 3.3.6.3-4. The recombinant hexokinase was shuttled into a pAG426GPD-EGFP Gateway destination vector following protocols in Section 3.3.6.5.

3.3.6.10 Lithium Acetate-mediated transformation of *Saccharomyces cerevisiae*

Saccharomyces cerevisiae cells were grown overnight in 5 ml YEPD at 30 °C with shaking at 180 rpm. 1 ml of the overnight culture was centrifuged at 4000 rpm for 5 minutes, washed once in water and resuspended in 1 M lithium acetate. 100 µl of the suspension was subsequently resuspended with premade transformation mix (240 µl PEG 3500 50 % w/v, 36 µl lithium acetate (LiAc) 1 M, 50 µl Boiled SS-carrier DNA, 34 µl vector) and incubated at 42 °C for 40 minutes. Cells were resuspended in 1 ml YEPD (YEPD/2 % galactose for YSH7.4-3C), left at 30 °C for an hour to recover. The suspension was subsequently plated on appropriate selection plates and incubated for 5 days.

3.3.6.11 Expression of microsporidian hexokinases in *Escherichia coli*

The hexokinase gene of *Ent. canceri* without its PTPA domain, the full hexokinase gene of *Ent. canceri*, *Enc. cuniculi* and *H. eriocheir*' hexokinase 2 were amplified by PCR with primers designed to adjunct Gateway recombination sites to the 5' and 3' ends of the hexokinase genes (See Appendix 3 for detailed description of primers). The hexokinase genes for *Ent. canceri* and *Enc. cuniculi* were amplified from genomic DNA whereas that of *H. eriocheir* was amplified from a commercially synthesized gene inserted into a pUC57 vector (Genwiz, UK). The amplified PCR product was gel purified using the Wizard SV kit (Promega, UK) and cloned into a pDONR221 vector following the Gateway

systems' BP cloning protocol (Gateway, UK). The resulting pDONR221-microsporidianHK plasmid was transformed into *E. coli* TOP10 cells (Invitrogen, UK) following transformation protocols in Section 3.3.6.6 and positive transformants were confirmed by colony PCR. Positive transformants were incubated overnight in 5 ml of Luria-Bertani (LB) broth containing 50 µg/ml Kanamycin and the propagated plasmids were harvested using a miniprep kit following manufacture's protocols (ThermoFisher Scientific, UK). The microsporidian hexokinase was shuttled from the harvested pDONR221-microsporidianHK plasmids into a pDEST17 *E. coli* expression vector following the Gateway systems LR cloning protocols (Gateway, UK). This resulted in a pEXP17-microsporidianHK plasmid. Competent *E. coli* TOP10 cells (Invitrogen, UK) were transformed with the cloned plasmid and positive transformants were confirmed by colony PCR using primers outlined in Appendix 3. Positive transformants were grown overnight in 5 ml of LB broth containing 100 µg/ml ampicillin and the propagated plasmids were harvested using a miniprep kit following manufacture's protocols (ThermoFisher Scientific, UK). The purified pEXP17-microsporidianHK plasmid was transformed into competent Rosetta(DE3)LysS cells (Invitrogen, UK) following protocols in Section 3.3.6.6 (See Appendix 2 for details of plasmids constructed in this study).

A single colony of the transformed cells was picked and inoculated in 5 ml of LB media, 100 µg/ml ampicillin. The overnight culture was centrifuged at 4000 rpm for 5 minutes in a precooled centrifuge and the supernatant media was carefully decanted. The pelleted cells were then transferred into a 1000 ml conical flask containing 400 ml LB media, 100 µg/ml ampicillin. The culture was incubated at 37 °C at 180 rpm until an optical density (OD₆₀₀) of 0.6 was reached. At this point, expression of the recombinant hexokinase protein, which was under the control of the lac operator, was induced by the addition of 1 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG). The culture was incubated at 30 °C at 180 rpm for 4 hours.

3.3.6.12 Expression of microsporidian hexokinases in *Saccharomyces cerevisiae*

In order to obtain recombinant microsporidian hexokinases for functional assays, the hexokinase of *Enc. cuniculi* was used for a pilot protein expression experiment in yeast cells. A poly HIS-tag was added to the C-terminal of the

translated protein and the gene was cloned into a PYES2 vector following protocols in 3.3.6.8. The resulting expression vector was transformed into a BY4741 yeast strain following protocols in Section 3.3.6.10. A single yeast colony confirmed by PCR to be a positive transformant was incubated in 1 L incomplete-URA3 SD/glucose overnight. The culture was centrifuged at 4000 rpm and the pellet was incubated in 1 L incomplete-URA3 galactose (not glucose) SD media overnight. Since the cloned hexokinase gene was under the control of GAL1 promoter, it was important for the overnight incubation to be carried out in SD/galactose media in order to induce heterologous expression of the recombinant protein. Subsequently, yeast cultures were centrifuged in a precooled ultracentrifuge at 3000 rpm for 5 minutes, rinsed in pre-chilled PBS and resuspended in pre-chilled lysis buffer containing NaH_2PO_4 50 mM pH 7.4, NaCl 250 mM and antiprotease cocktail (ThermoFisher Scientific, UK). Yeast lysates were centrifuged at 13000 rpm for 15 minutes and the supernatant was used for Dynabeads HIS-tag protein isolation (Novex, UK) following manufacturer's manual. Purified proteins were resolved by running them on a 10%-20% SDS-PAGE gel (sodium dodecyl sulphate polyacrylamide gel electrophoresis) (ThermoFisher Scientific, UK) under reducing conditions. The protein gel was stained using SimplyBlue (Invitrogen, UK) and processed for mass spectrometry analysis.

For functional complementation experiments, YSH7.4-3C, a hexokinase triple knockout strain was transformed with *Ent. canceri*'s chimeric hexokinase cloned into Gateway's pAG426GPD-EGFP destination vector. The benefit of this destination vector was that it juxtaposed a green fluorescent protein (GFP) tag at the N-terminus of the translated recombinant protein thereby allowing detection of subcellular localization by fluorescence microscopy. The cloning and transformation protocols were as per Section 3.3.6.9 and 3.3.6.10. Growth of transformed cells was recorded with the G:Box imager (Syngene, UK) and fluorescence and bright field microscopy was performed on individual cells an Olympus Ix81 Inverted fluorescent microscope (University of Exeter Bioimaging Centre, UK).

3.3.6.13 Recombinant protein HIS-tag purification

IPTG induced cells were harvested by centrifugation at 4000 rpm for 5 minutes in a precooled centrifuge. Pelleted cells were resuspended in pre-chilled 700 µl washing/binding buffer (Section 3.3.5.3) and lysed with a sonicator (Bio-rad, UK). Sonication was performed 5 times at full power for 30 seconds interrupted by a 2 minute cooling period to ensure lysis of all cells but also to minimize protein degradation resulting from high temperature. The lysed cells were centrifuged at 13000 rpm for 15 minutes. 500 µl of the supernatant was used for subsequent His-tag purification steps. His-tag purification was performed following the Novex His-tag Dynabead purification protocol.

3.3.6.14 Western blotting

Purified proteins were resolved by running them on a 10%-20% SDS-PAGE gel under reducing conditions. Western blotting was performed by electroblotting protein bands onto a Polyvinylidene difluoride (PVDF) membrane (ThermoFisher Scientific, UK) for 2 hours at 30 V. This was performed in ice cold conditions. Following this, the electroblotted PVDF membrane was blocked for an hour at room temperature in blocking buffer [PBS-Tween-20 (0.01 %), 5 % non-fat milk] and subsequently washed for five minutes in PBS-Tween-20 (0.01 %). The membrane was then incubated in primary antibody buffer [(1:2000) Polyclonal anti-His mouse antibodies (ThermoFisher Scientific, UK) in PBS-Tween-20 (0.01 %) and 1 % milk] at 4 °C overnight with gentle rocking. The membrane was then washed five times for 10 minutes in PBS-Tween-20 (0.01 %). Secondary antibody incubation was performed by soaking the membrane in secondary antibody buffer [(1:5000) HRP-conjugated anti-mouse goat antibodies (Sigma, UK) in PBS-Tween-20 (0.1 %), 1 % milk] for an hour with gentle rocking at room temperature. The membrane was then washed five times for 5 minutes in PBS-Tween-20 (0.1 %). The final two washes were performed in PBS. Pierce ECL kit (ThermoFisher Scientific, UK) was used for signal development by following manufacturer's instructions. Excess reagent was removed and image acquisition was performed using darkroom developing techniques and a G:Box imager (Syngene, UK).

3.3.6.15 Construction of an empty pEXP17 clone for use in BIOLOG analysis

For BIOLOG analysis, it was important to use Rosetta cells transformed with an empty pEXP17 vector as a negative control. However, an empty pDEST17 vector

contains the *ccdB* gene that ultimately kills transformed cells. EcoRI restriction enzyme sites located on the vector were used to remove the coding DNA sequence of the *ccdB* gene. Volumes of the EcoRI enzyme (New England BioLabs, UK) and buffers used were as per manufacturer's instructions. The T4 DNA ligase (New England BioLabs, UK) was used to ligate the sticky ends of the digested plasmid as per manufacturer's instructions See Appendix 2 for detailed map of empty pEXP17 vector.

3.3.6.16 BIOLOG Analysis

Rosetta2(DE3)pLYS *E. coli* cells (Millipore, UK) transformed with pDEST17 vectors (Gateway, UK) containing hexokinases from *Enc. cuniculi*, *Ent. canceri*, *H. eriocheir* hexokinase 2 and an empty vector were submitted to the BIOLOG facility at the University of Exeter. The cells were growing on LB agar, 100 µg/ml ampicillin at the time they were submitted. Here, equal concentrations of the transformed bacterial cells were suspended in redox dye Mix G (BIOLOG, UK) complemented with 100 µg/ml ampicillin to maintain the integrity of the transformed plasmid and 1mM IPTG to induce expression of the recombinant hexokinase. The *E. coli* suspension for each microsporidian hexokinase was dispensed onto a 95-carbon source microplate PM-1 (BIOLOG, UK). A microplate reader was used to monitor oxidation of the various carbon substrates by the transformed bacteria every 15 minutes for 72 hour at 25 °C.

3.3.6.17 Hexokinase activity assay

Activity assays were conducted for *Ent. canceri*'s and *H. eriocheir*'s HIS-tag purified recombinant hexokinases following modified protocols from Claeysen *et al.* 2006. Here, the 200 µl reaction mix contained 1.4 U/ml glucose-6-phosphate dehydrogenase (G2921 Sigma, UK), 50 mM Tris-HCL, pH 8.0, 50 mM KCL, 5mM MgCl₂, 5 mM DTT, 0.3 mM NAD⁺, 1mM ATP, 5mM glucose and 10 ng/ml recombinant hexokinase. Similar reactions with varying hexokinase concentrations (5 ng/ml, 2 ng/ml and 1 ng/ml) were also set on the same microwell plate. The positive control contained *Saccharomyces cerevisiae*'s hexokinase (H4502 Sigma, UK). The negative control contained all components of the reaction apart from hexokinase; Instead, elution buffer was used to replace hexokinase volumes. Reactions were performed at the same time in a 96-well Polystyrene microwell plate (Nunc™ ThermoFisher Scientific, UK). The reaction was incubated at 30 °C and hexokinase activity was monitored by taking

absorbance readings of NADH at 340 nm every 5 seconds. The SpectraMax Microplate Reader used here was set to shake the plate for 2 seconds prior to taking each reading. Three repeats were set for assays performed with *Ent. canceri*'s recombinant protein whereas only two repeats were set for those performed with *S. cerevisiae* and *H. eriocheir* hexokinase. In order to remove background noise, absorbance readings from the negative control were subtracted from those performed with hexokinase.

3.3.6.18 Processing of protein samples for Matrix-assisted laser

Desorption/Ionization Mass Spectrometry (MALDI-MS) analysis

HIS-tagged purified proteins expressed in BY4741 yeast strains were resolved by running them on a precast 10%-20% SDS-PAGE gel (ThermoFisher Scientific, UK) under reduced conditions. Protein bands in the molecular weight region of 50 and 25 kDa were excised from the protein gel. Bands in the range of 50 kDa were selected because that is the expected *Enc. cuniculi* hexokinase protein size whereas bands at 25 kDa were selected because they were highly expressed. The protein gel was washed 3 times for 5 minutes with 100 ml deionized water to remove residual SDS and buffer salts, which interfere with binding of the dye to the protein. The gel was submerged in SimplyBlue dye (Invitrogen, UK) and left at room temperature with gentle shaking for 5 minutes. The gel was subsequently rinsed in deionized water for 1-3 hours. The bands of interest were excised from the gel by using a sterile scalpel and transferred into sterile microcentrifuge tubes. The excised gels were destained by adding 300 µl of destaining buffer (50 % acetonitrile in 50 mM ammonium bicarbonate) and incubated for 30 minutes at 30 °C with periodic vortexing. Dehydration of the gel was performed by carefully removing the destaining buffer and submerging the excised gels in 200 µl of acetonitrile. The gels were rehydrated by the addition of 200 µl of rehydration buffer [10 mM Dithiothreitol (DTT) in 50 mM ammonium bicarbonate] and incubated at 50 °C for 45 minutes. DTT was subsequently removed by adding 200 µl 50 mM iodoacetamide to the gel and left in the dark for 45 minutes at room temperature.

The gel band was transferred into a clean microcentrifuge tube and immersed in 300 µl destaining buffer for 2-3 minutes. This was repeated thrice. The gel bands were further cut into smaller pieces with a sterile scalpel and centrifuged at 13000 rpm for 2 minutes. After removing the supernatant, the pelleted gel pieces were

covered with 200 μ l acetonitrile and incubated at room temperature until all the gel pieces turned completely white in colour. Acetonitrile was carefully pipetted out of the microcentrifuge tubes and gel pieces were left to dry in the hood. Tubes were subsequently stood on ice and gel pieces were rehydrated by adding 60 μ l of porcine trypsin (Promega, UK). Samples were left on ice for 15 minutes. Gel pieces were covered with 100 μ l of ammonium bicarbonate and the cap of the microcentrifuge tubes were sealed with Parafilm (Sigma, UK). Samples were incubated overnight at 37 °C. The next day samples were sonicated for 15 minutes and centrifuged at 13000 rpm for 15 minutes. The supernatant was decanted into a clean microcentrifuge tube and centrifuged again for 15 minutes at 13000 rpm. The top 20 μ l of the supernatant was sent off for MALDI-MS analysis at the University of Exeter's Mass Spectrometry Facility.

3.4 Results

3.4.1 Loss of glycolysis in the Enterocytozoonidae

After performing orthology clustering of predicted proteins encoded by the genomes sequenced in this study with those of publicly available genomes it was evident that the genomes of member species of the Enterocytozoonidae family lacked a complete set of proteins for several pathways (summarized in Figure 3.2), including glycolysis. The genomes of *Ent. canceri*, *E. bieneusi*, *E. hepatopenaei* and *Hepatospora* spp. did not encode for all 10 glycolytic proteins whereas that of the other 18 non-Enterocytozoonidae microsporidian genomes encoded for all glycolytic proteins (Figure 3.3). Different species within the Enterocytozoonidae lacked different glycolytic proteins although both *Hepatospora* sub-species displayed the same pattern of loss (Figure 3.3). Four glycolytic proteins were absent in all Enterocytozoonidae genomes, namely Phosphofructokinase, Fructose-bisphosphate aldolase, Phosphoglycerate kinase and Pyruvate kinase. Hexokinase was the only protein encoded by all genomes analysed in this study (Figure 3.3).

Apart from proteins directly involved in metabolism, members of the Enterocytozoonidae lacked certain proteins needed for redox balance of NADH and NAD⁺ such as AOX (Figure 3.4). Absence of this protein was however not exclusive to Enterocytozoonidae genomes. Interestingly, a scan of orthologous protein clusters exclusive to glycolytic genomes yielded only 38 clusters, many of which did not have a predicted function (Appendix 14). Four of these clusters were the glycolytic proteins absent in the Enterocytozoonidae mentioned above. One cluster was for the bis(5'-nucleosyl)-tetraphosphatase enzyme family also known as nucleoside diphosphate linked to some moiety, X (NUDIX) hydrolases. The majority of the remaining protein clusters were enzyme families belonging to the mevalonate pathway (Figure 3.5)

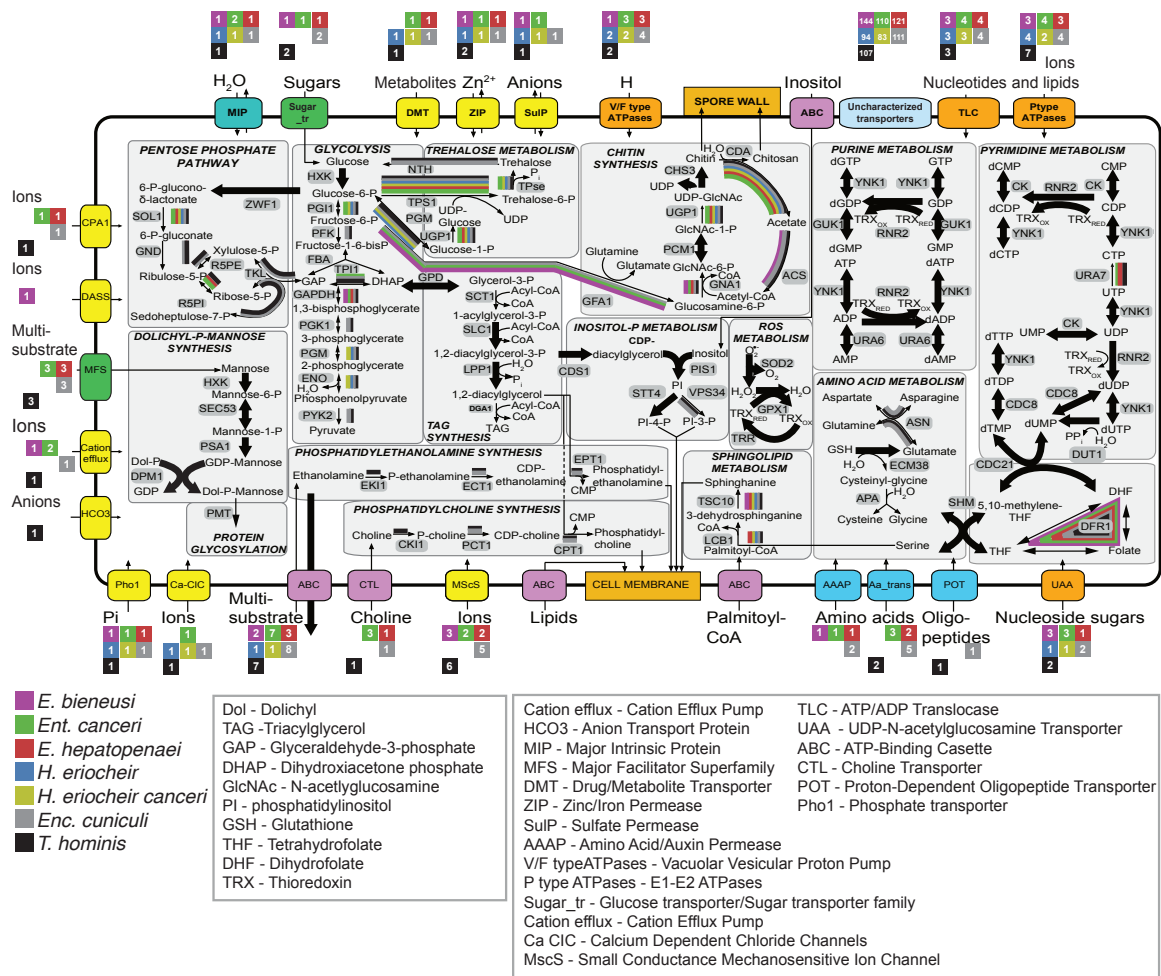


Figure 3.2: Summary of metabolic pathways retained in the Enterocytozoonidae as revealed by comparative genomic survey. Plasma membrane transporter families with their respective copy numbers in each of the analyzed genomes are also displayed. Enterocytozoonidae genomes sequenced in this study were compared to that of *E. bieneusi* and 19 other non-Enterocytozoonidae genomes. However, results for only *Enc. cuciculi* and *T. hominis* were displayed here. From Wiredu-Boakye *et al.* submitted, see appendix 13)

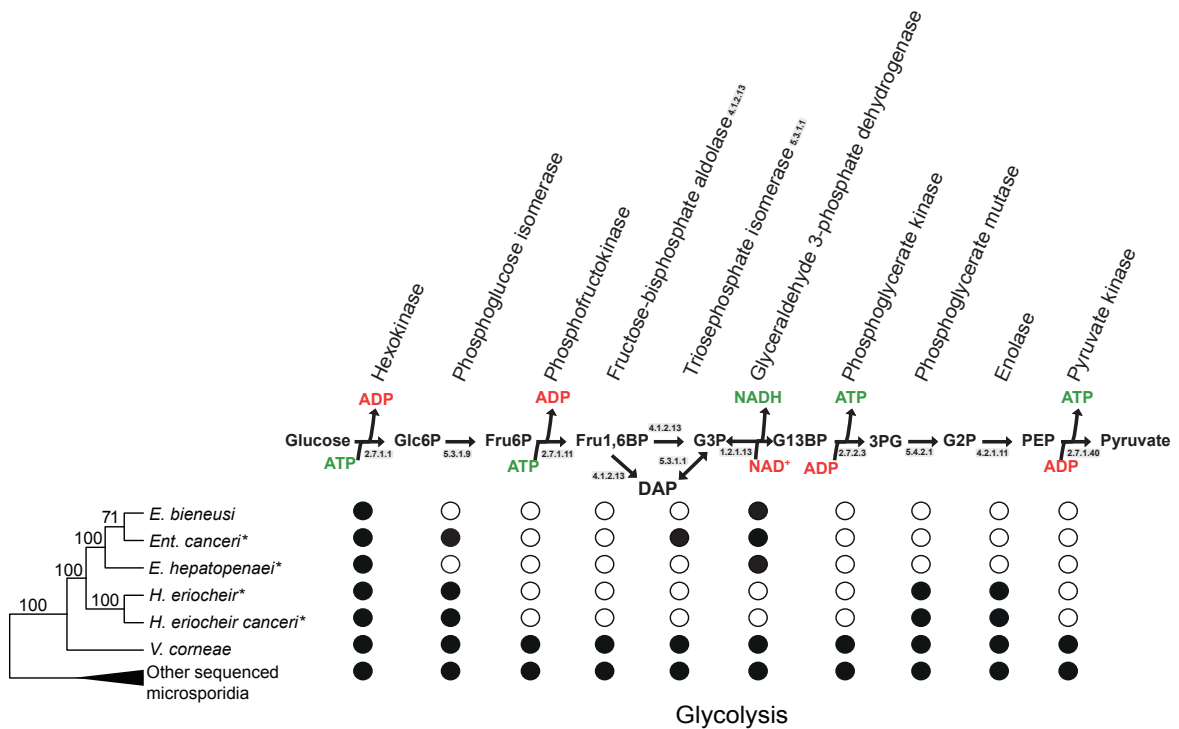


Figure 3.3: Loss of glycolytic genes within the Enterocytozoonidae. Arrows symbolize the substrate flux of the enzymatic reactions. The release or breakdown of ATP or reducing potential in the form of NADH is indicated in green or red respectively. Enzymes that catalyse various reactions are indicated on top of each arrow with their EC numbers indicated on grey background. Black circles represent presence of gene whereas white circles represent absence. Abbreviations: PEP, phosphoenolpyruvate; DAP, dihydroxyacetone phosphate; Fru6P, fructose 6-phosphate; Fru1,6BP, fructose 1,6-bisphosphate; Glc6P, glucose 6-phosphate; G3P, 3-phosphoglycerate; G13BP, 1,3-bisphosphoglycerate; G2P, 2-phosphoglycerate. (Adapted from from Wiredu-Boakye *et al.* submitted, see appendix 13).* organisms sequenced in this study.

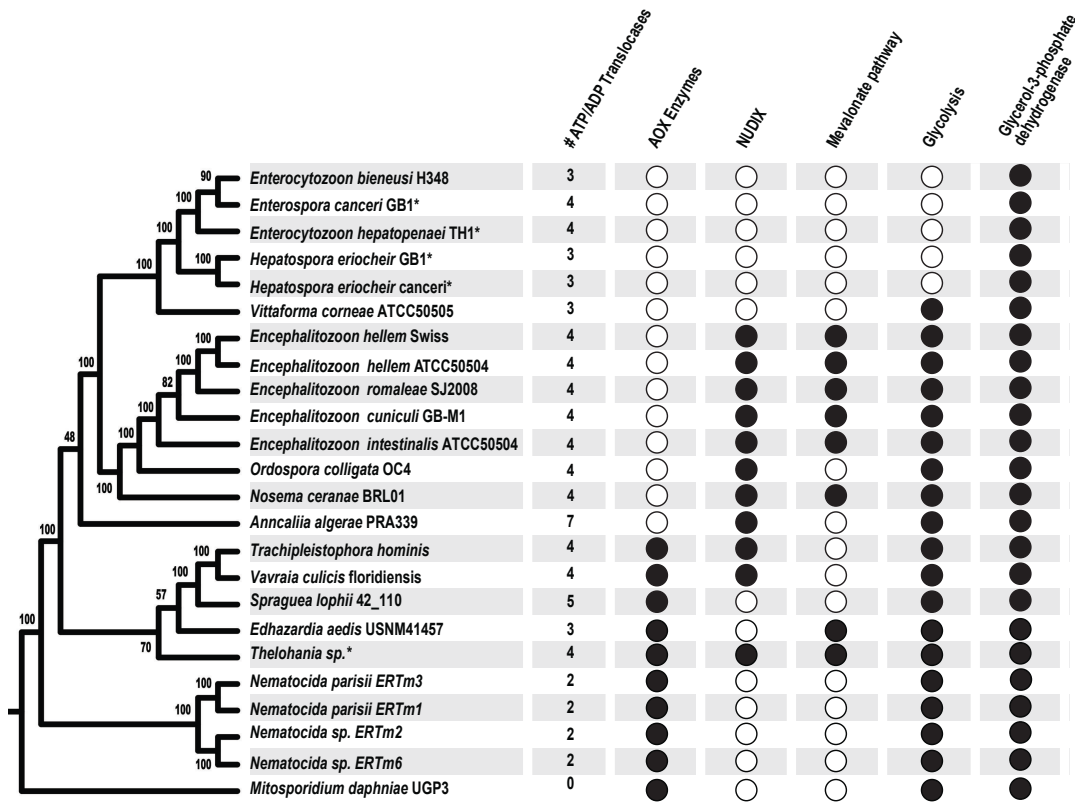


Figure 3.4: Explaining loss of glycolysis in the Microsporidia by mapping ATP/ADP translocase numbers present in microsporidian genomes and NAD⁺ replenishing pathways/enzymes onto the microsporidian phylum. * organisms sequenced in this study.

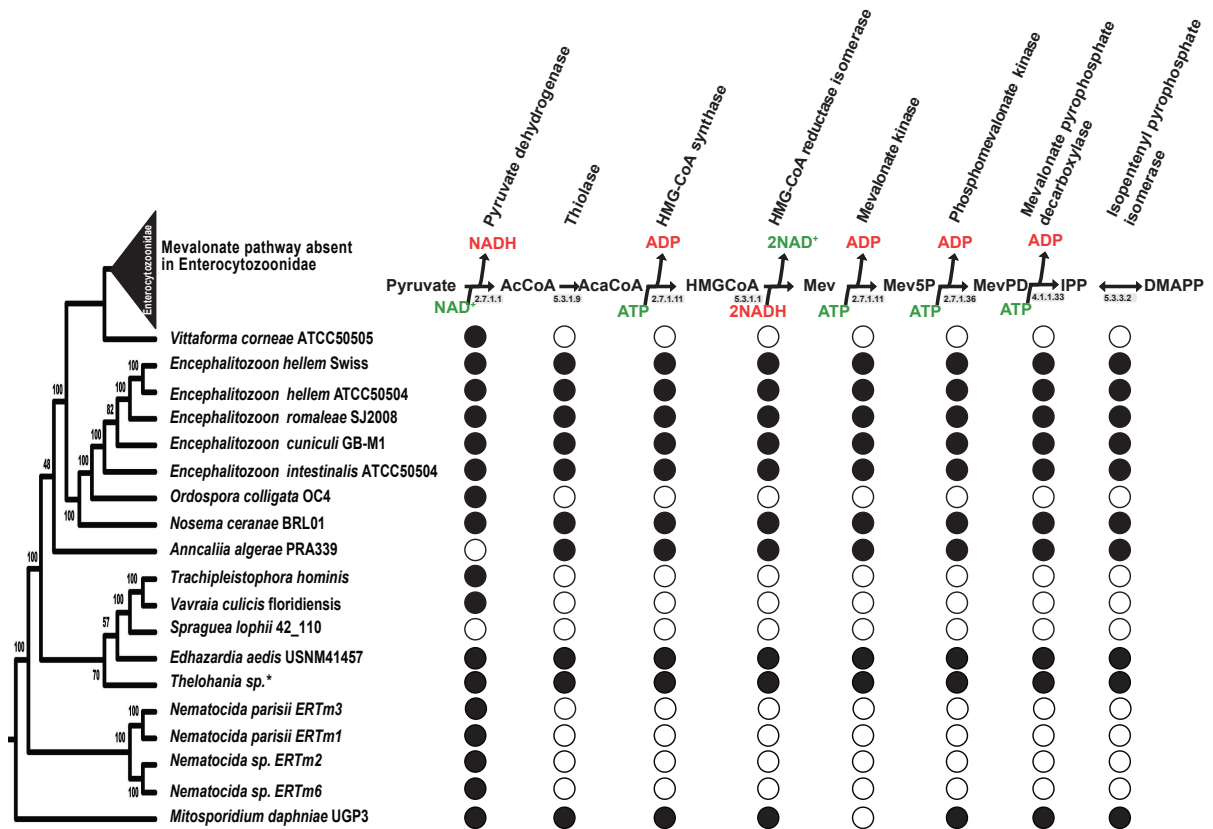


Figure 3.5: Linking glycolysis to the mevalonate pathway: Genes involved in the mevalonate pathway and their distribution in the microsporidian phylum. The release or consumption of energy or reducing potential is indicated in green or red respectively. * organisms sequenced in this study.

3.4.2 Phylogenetic assessment of microsporidian hexokinases

In this analysis hexokinases and glucokinases of *H. sapiens*, *R. rattus* and *S. cerevisiae* branched at the base of the tree. There was strong statistical support for the branching of the entire microsporidian hexokinase clade from glucokinase of *S. cerevisiae* and the single hexokinase of *R. allomycis*. The hexokinase from *M. daphniae* branched at the base of the microsporidian clade. Individual clades such as *Nematocida*, *Vavraia/Trachipleistophora*, *Nosema*, *Encephalitozoon* were well supported statistically. Some microsporidian lineages possessed multiple hexokinases which may have stemmed from recent lineage-specific duplications. More specifically, there appeared to have been multiple duplications in the genome of *E. bienewisi*, a single duplication in the genome of *Vittaforma*, a single duplication in the common ancestor of *Trachipleistophora* and *Vavraia*, and in the common ancestor of the two *Hepatospora* species (Figure 3.6 and 3.7). An unexpected observation was the grouping of *Vittaforma* hexokinases with those of *Encephalitozoon* clade with strong statistical support (Figure 3.6 and 3.7). Even after the manual removal of the PTPA domain from the protein sequence

of *Ent. canceri*'s hexokinase, its position on ML trees remained the same (Figure 3.7).

To assess whether the microsporidian hexokinases are indeed functional as glycolytic enzymes, active sites (oligopeptide stretches and single amino acids) that have been empirically shown to be essential for the normal function of *S. cerevisiae*'s hexokinase 1 were mapped onto the microsporidian hexokinases. Active sites of *M. daphniae* and early branching species such as *Nematocida* spp. were highly similar to those of *S. cerevisiae*. An interesting observation was that in instances where microsporidian genomes encoded for more than one hexokinase, the active sites in one of the protein copies was less conserved (See *T. hominis*, *N. bombycis* and *E. bieneusi* in Figure 3.6). Furthermore, hexokinases used in this study were assessed for the possession of signal peptides at their N-terminal. The acquisition of a signal peptide appeared to be species specific as no noticeable pattern of acquisition was observed. Hexokinases of early branching *Nematocida* spp. and human parasites *Enc. cuniculi* GB and *Enc. intestinalis* possessed signal peptides. Interestingly, hexokinases of all other *Encephalitozoon* substrains and species did not possess signal peptides. Similarly, the hexokinase of *N. ceranae* possessed a signal peptide but those of all other *Nosema* species did not (Figure 3.6). A hexokinase encoded by the genome of *Vav. culicis* and two homologs of *T. hominis* also possessed a signal peptide. All non-microsporidian hexokinases analysed did not possess a signal peptide. The most unexpected observation from this analysis was that the N-terminus of the *Ent. canceri* hexokinase encoded for a phosphotyrosyl phosphate activator (PTPA) domain.

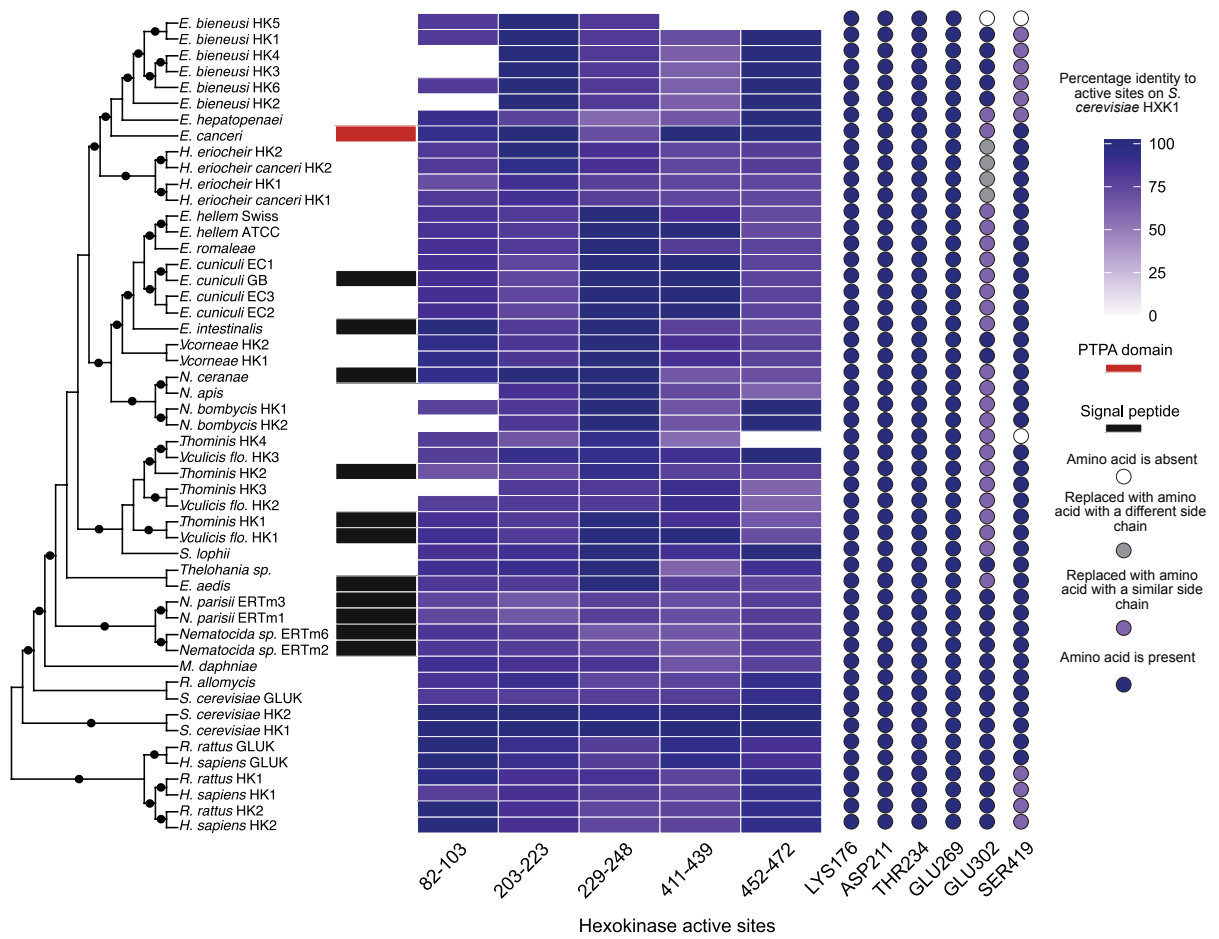


Figure 3.6: Phylogeny of hexokinase reveals that it is duplicated in some microsporidian lineages: The phylogenetic positions are derived from maximum likelihood analysis performed on a masked alignment of hexokinase proteins. Mammalian hexokinases were used to root the tree. Support values that are above 75 % have been represented by black dots on the respective nodes. Similarity between *S. cerevisiae*'s hexokinase active sites and those of other lineages was represented with a heat map. (From Wiredu-Boaky *et al.* submitted, see appendix 13)

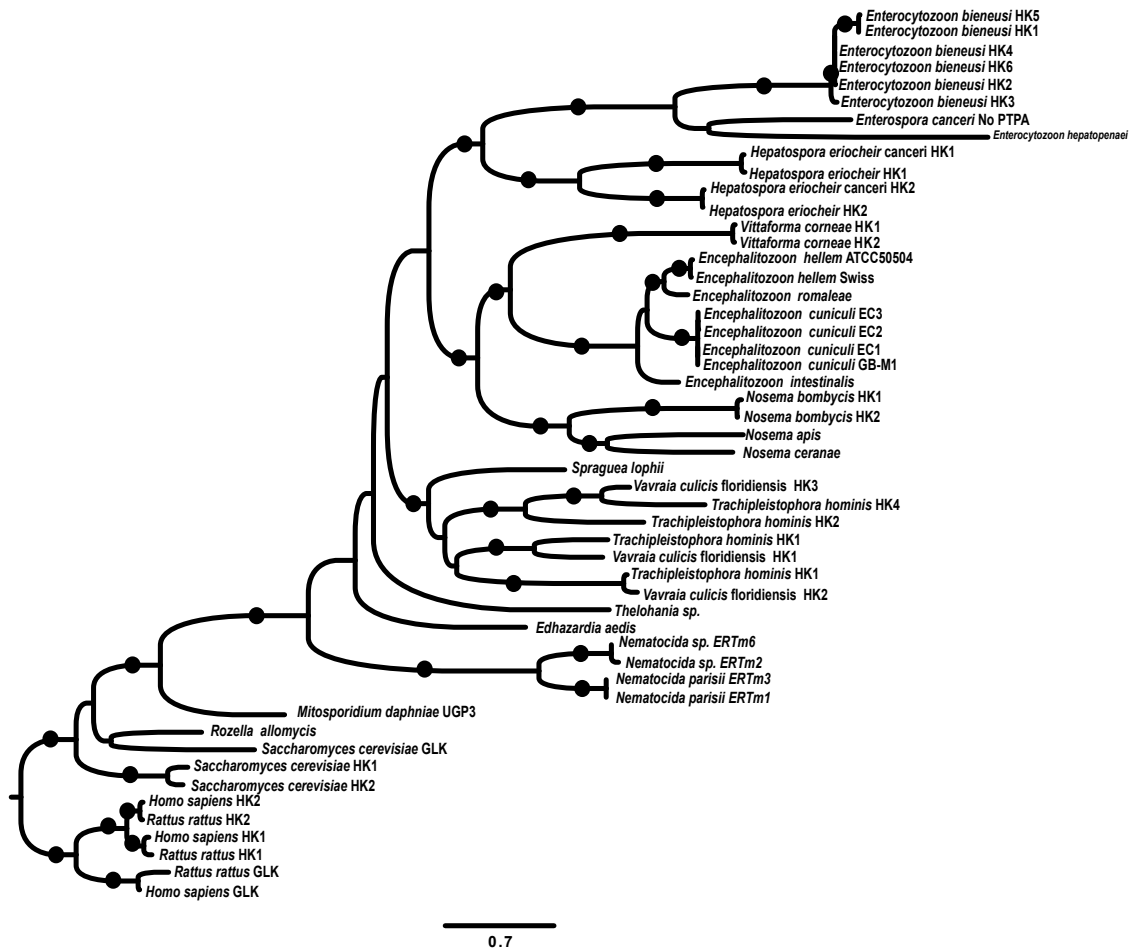


Figure 3.7: Maximum likelihood analysis performed on a masked alignment of microsporidian hexokinases (HK). *Ent. canceri*'s PTPA domain fused to the N-terminal of its hexokinase was manually deleted prior to the alignment: Bootstrap inferences above 75 % have been represented by black dots on the respective branches. Human and rat hexokinase and glucokinase (GLK) were used to root the tree. The scale represents number of amino acid substitutions per site.

3.4.3 Phylogenetic assessment of microsporidian PTPA proteins

The two homologs from the *M. daphniae* branched at the base of the microsporidian phylum. Homologs from clades such as *Nematocida*, *Encephalitozoon* and *Vavraia/Trachipleistophora* branched together. The PTPA homolog of the intranuclear crab parasite, *Ent. canceri* was the closest relative of those of the human parasite, *E. bieneusi*. All 12 PTPA homologs of *E. bieneusi* clustered together with high bootstrap support. Overall, these phylogenetic inferences were well supported by bootstrap analysis (Figure 3.8).

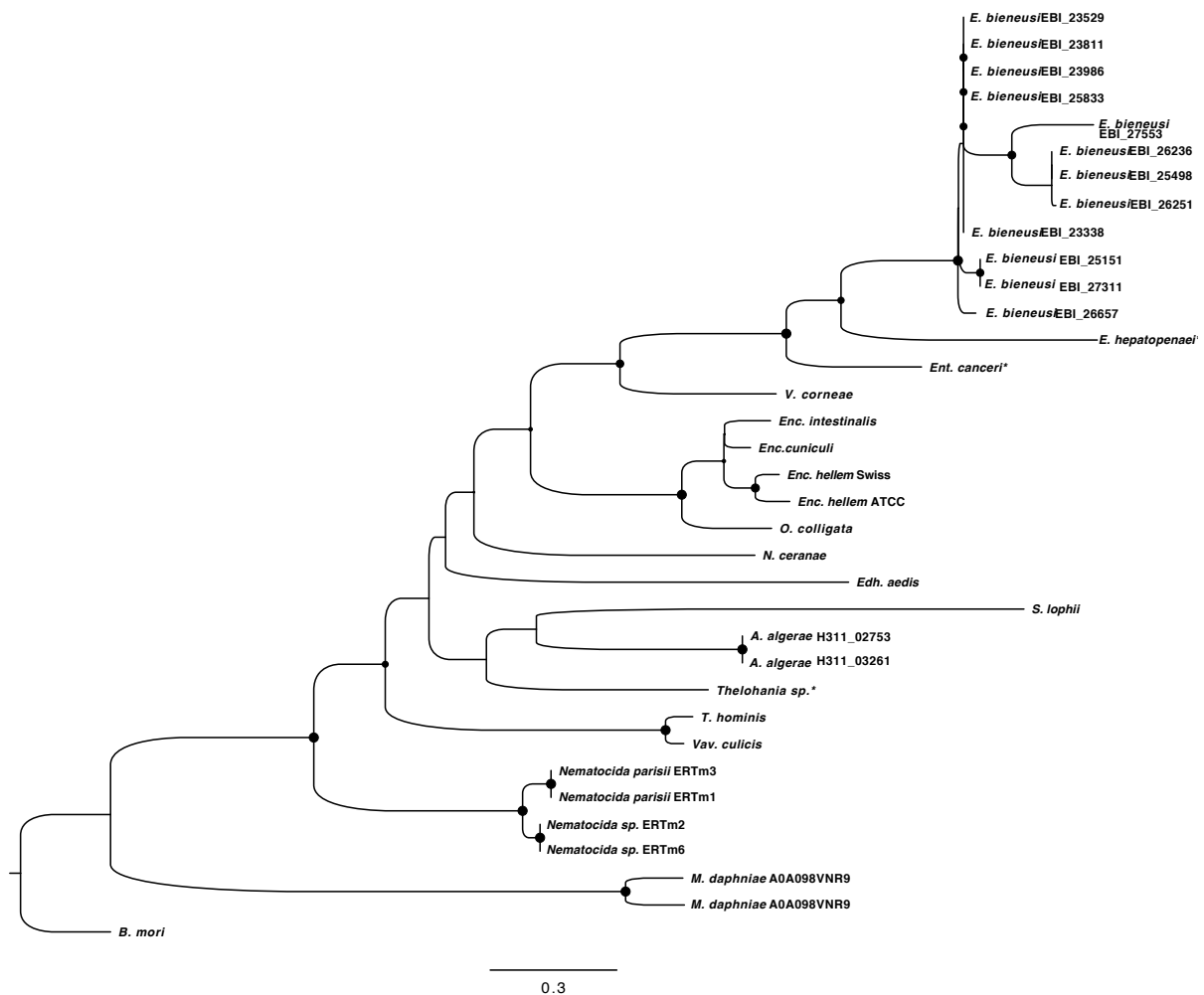


Figure 3.8: Maximum likelihood analysis performed on a masked alignment of microsporidian PTPA proteins. *Ent. canceri*'s PTPA domain fused to the N-terminal of its hexokinase was used in this analysis: 100 bootstrap inferences above 70 are displayed on their respective nodes as circles. The scale represents number of amino acid substitutions per site. The silkworm, *Bombyx mori* homolog of the protein was used to root the tree.

3.4.4 Evolution of the chimeric hexokinase in *Enterospora canceri*

To investigate the evolutionary processes that led to the *Ent. canceri* PTPA-hexokinase chimera, gene order conservation of contigs containing hexokinases and PTPAs across members of the Enterocytozoonidae and *V. corneae* were compared (Figure 3.9-11). *E. bienewisi* was omitted from these analyses as it possessed multiple numbers (>5) of hexokinases and PTPA homologs scattered across multiple short contigs. Inclusion of these contigs would have made graphical representation of the data difficult. *V. corneae* is the closest phylogenetically related species of the Enterocytozoonidae whose genome is publicly available. Assessment of gene order conservation between *V. corneae* and members of the Enterocytozoonidae show that hexokinase and PTPA are in genetic vicinity only in the genomes of *Ent. canceri* and *E. hepatopenaei*. Gene

order was highly conserved between *Ent. canceri* and *E. hepatopenaei* and less so between these two species and *H. eriocheir*. Gene order appeared to be more conserved between *E. hepatopenaei/Ent. canceri* and *V. corneae* than between *E. hepatopenaei/Ent. canceri* and *H. eriocheir* (Figure 3.9 and 3.10).

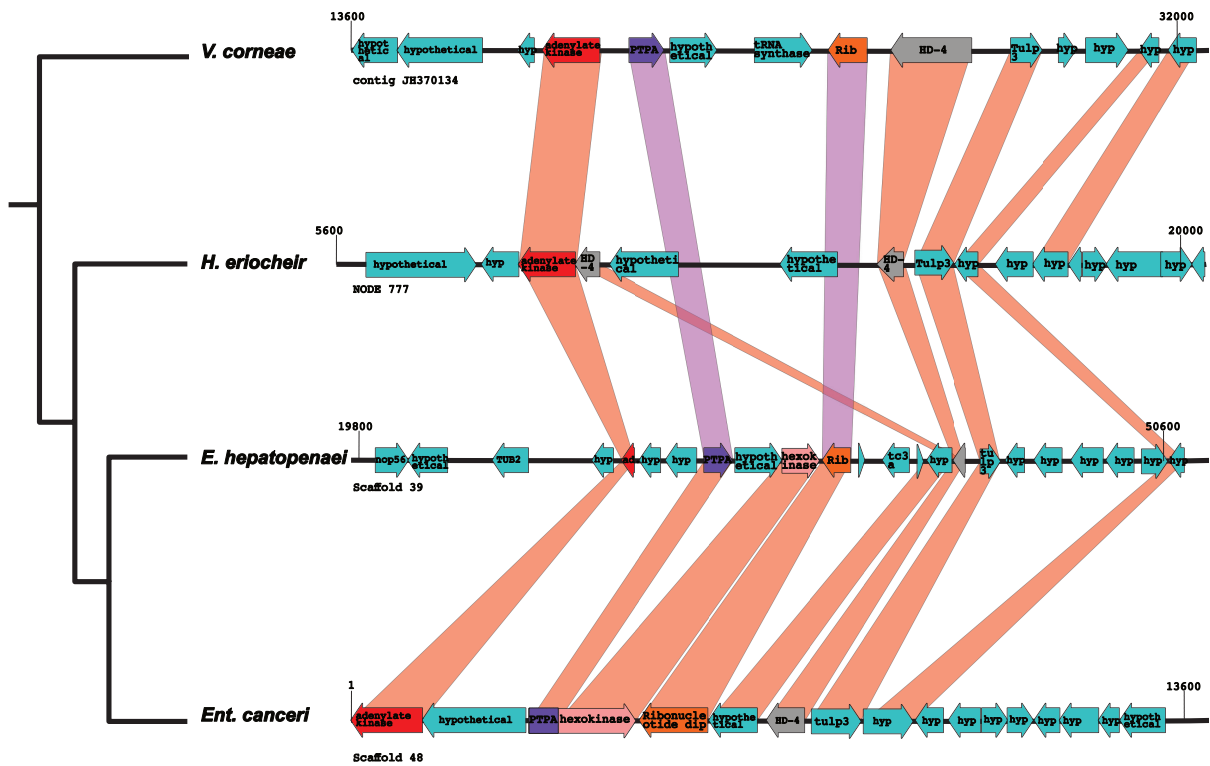


Figure 3.9: Comparing gene order conservation between hexokinase-containing contig in *Enterospora canceri* and corresponding contigs within the *Enterocytozoon hepatopenaei*, *Hepatospora eriocheir* and *Vittaforma corneae*. Coordinates of genes were determined with the ARTEMIS genome viewer but the maps were manually constructed in AFFINITY DESIGNER by performing independent BLASTP searches. The hexokinase ORF has been coloured pink whereas the PTPA ORF has been coloured purple. Adenylate kinase, ribonucleotide diphosphate and HD-4 were coloured red, orange and gray respectively. Other ORFs were coloured blue. Genes present in *V. corneae*, absent in *H. eriocheir* but present in any of the other two species have been mapped with purple for easy visualization.

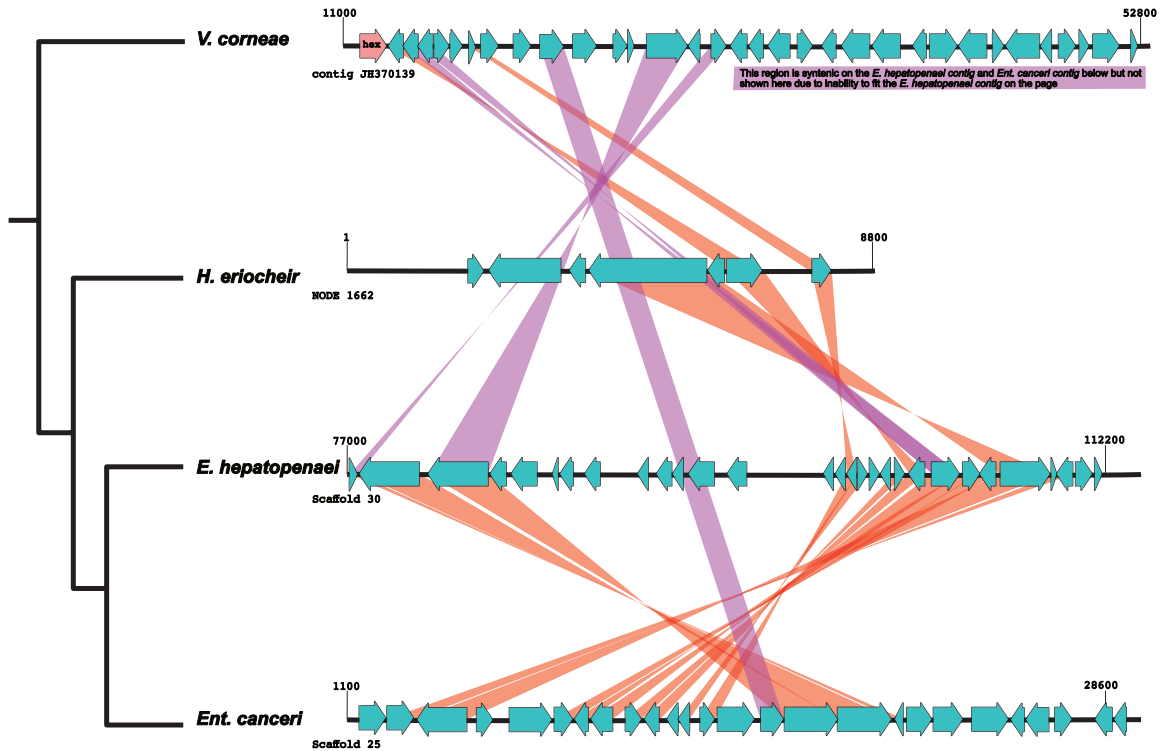


Figure 3.10: Comparing gene order conservation between hexokinase-containing contig in *Vittaforma corneae* and corresponding contigs within the Enterocytozoonidae. Coordinates of genes were determined with the ARTEMIS genome viewer but the maps were manually constructed in AFFINITY DESIGNER by performing independent BLASTP searches. The hexokinase ORF has been coloured pink whereas other genes are blue. Genes in *V. corneae*, absent in *H. eriocheir* but present in any of the other two species have been mapped with purple for easy visualization.

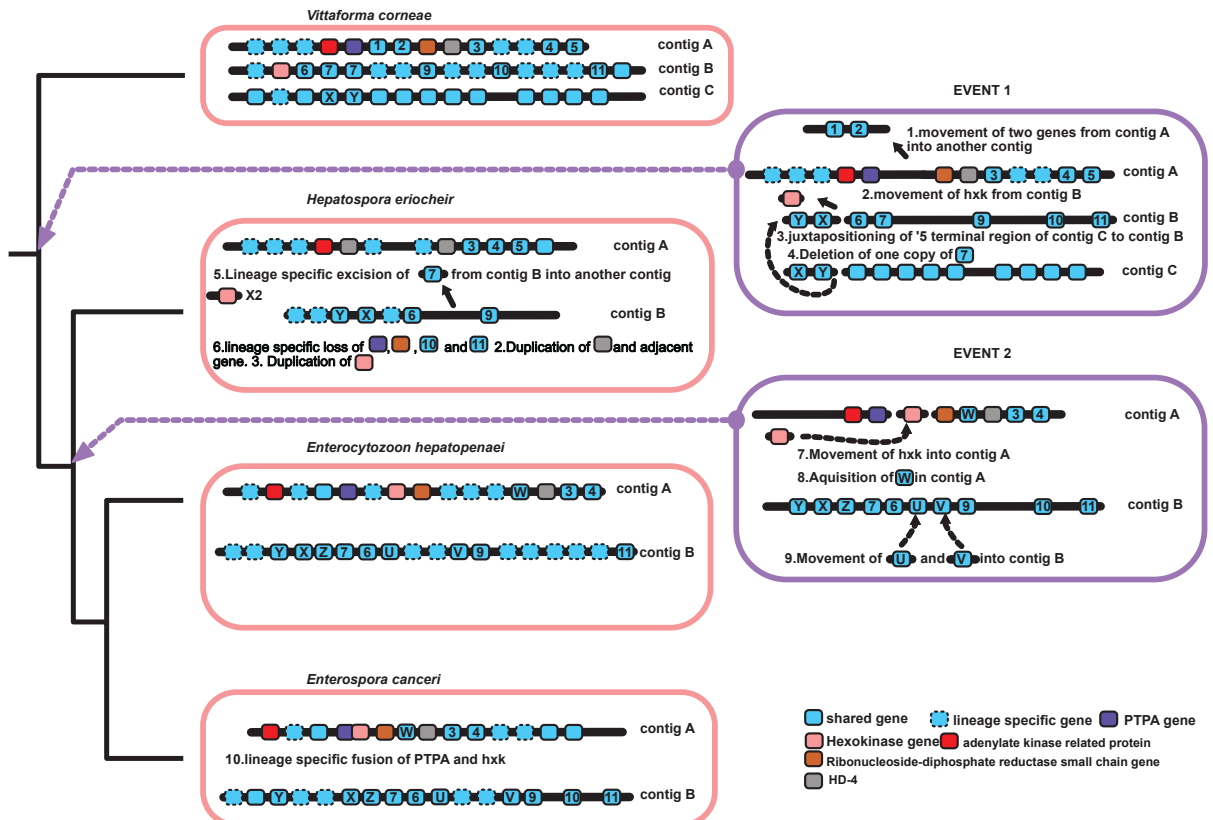


Figure 3.11: Understanding the evolution of Chimeric hexokinase in *Enterospora canceri*

3.4.5 Genome wide analysis of chimeric genes in the Microsporidia

Following the discovery of a chimeric hexokinase in *Ent. canceri*, genomes of all microsporidians analysed here were parsed for the presence of chimeric proteins to assess if formation of chimeric proteins was a common phenomenon in this phylum. The genome of *M. daphniae* (the outgroup) by far contained the highest number of chimeric proteins, 10. The other chimeric proteins identified in this analysis were found in the genomes of *N. ceranae*, *A. algerae*, *T. hominis*, *Vav. culicis*, *S. lophii*, *Edh. aedis*, *Thelohania* spp., *Nematocida* spp. and *Ent. canceri* (Figure 3.12). None of the genomes of *Encephalitozoon* species encoded for chimeric proteins. Apart from *Ent. canceri*, none of the Enterocytozoonidae genomes encoded for chimeric proteins (Figure 3.12).

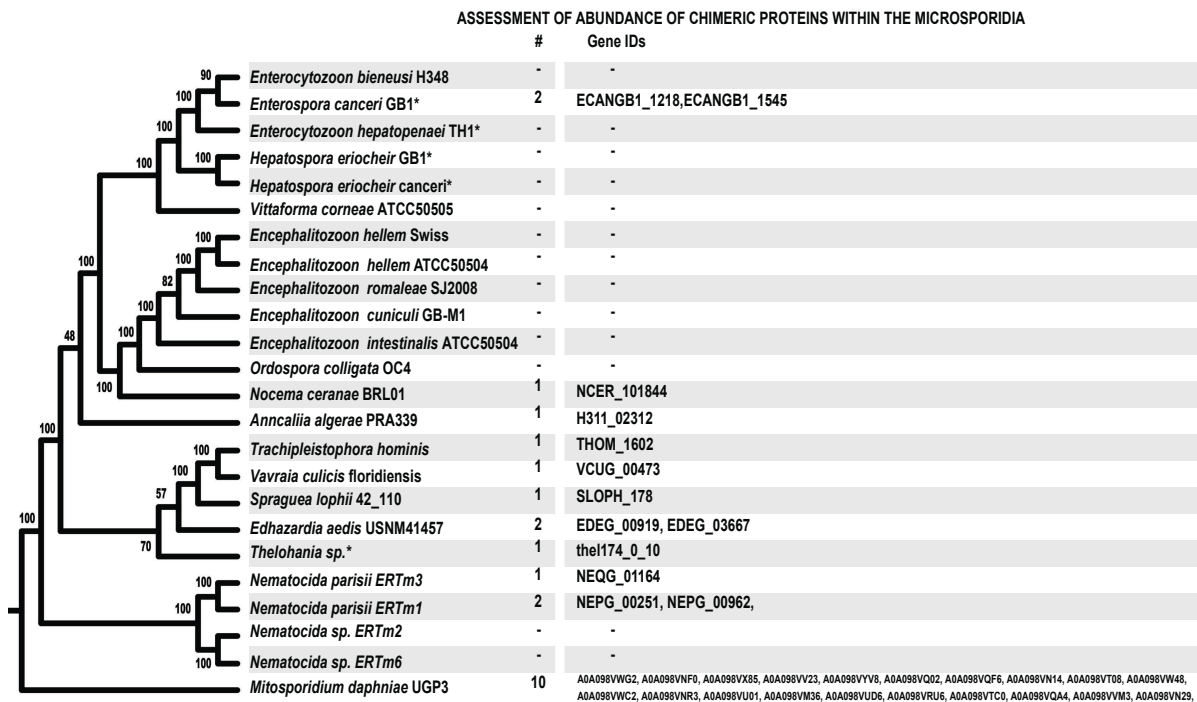


Figure 3.12: Putative chimeric proteins in the genome of extant microsporidians.

3.4.6 Phylogenetic assessment of microsporidian ATP/ADP translocases

In an effort to answer the question of how aglycolytic microsporidian lineages obtain ATP, a phylogenetic assessment of ATP/ADP translocases encoded by genomes of these lineages was performed. No homolog for this protein was observed for the genome of *Mitosporidium daphniae*. Although microsporidian homologs were very divergent, there was strong statistical support for the branching of microsporidian homologs as a sister clade of ATP/ADP translocases

belonging to *Rickettsia/Chlamydia/Arabidopsis* species. Homologs of early branching lineages such as *Nematocida* spp., *Anncalia algarae*, *Trachipleistophora hominis*, *Spraguea lophii* and *Vavraia culicis* formed separate clades, independent of the four clades containing homologs from *Encephalitozoon* spp., *Nosema* spp. and *Ordospora colligata*. All of the ATP/ADP translocases belonging to members of the Enterocytozoonidae could be found within one of these clades. However this clade could be further split into two subclades with one of the subclades accommodating only homologs of the Enterocytozoonidae member species and *V. corneae*. The second subclade contained single homologs of member species of the Enterocytozoonidae together with those from *V. corneae*, *Encephalitozoon* spp., *Nosema* spp. and *Ordospora colligata*. There was strong statistical support at the base of these clades. However, the phylogenetic relationship between individual clades was not strongly supported (Figure 3.13).

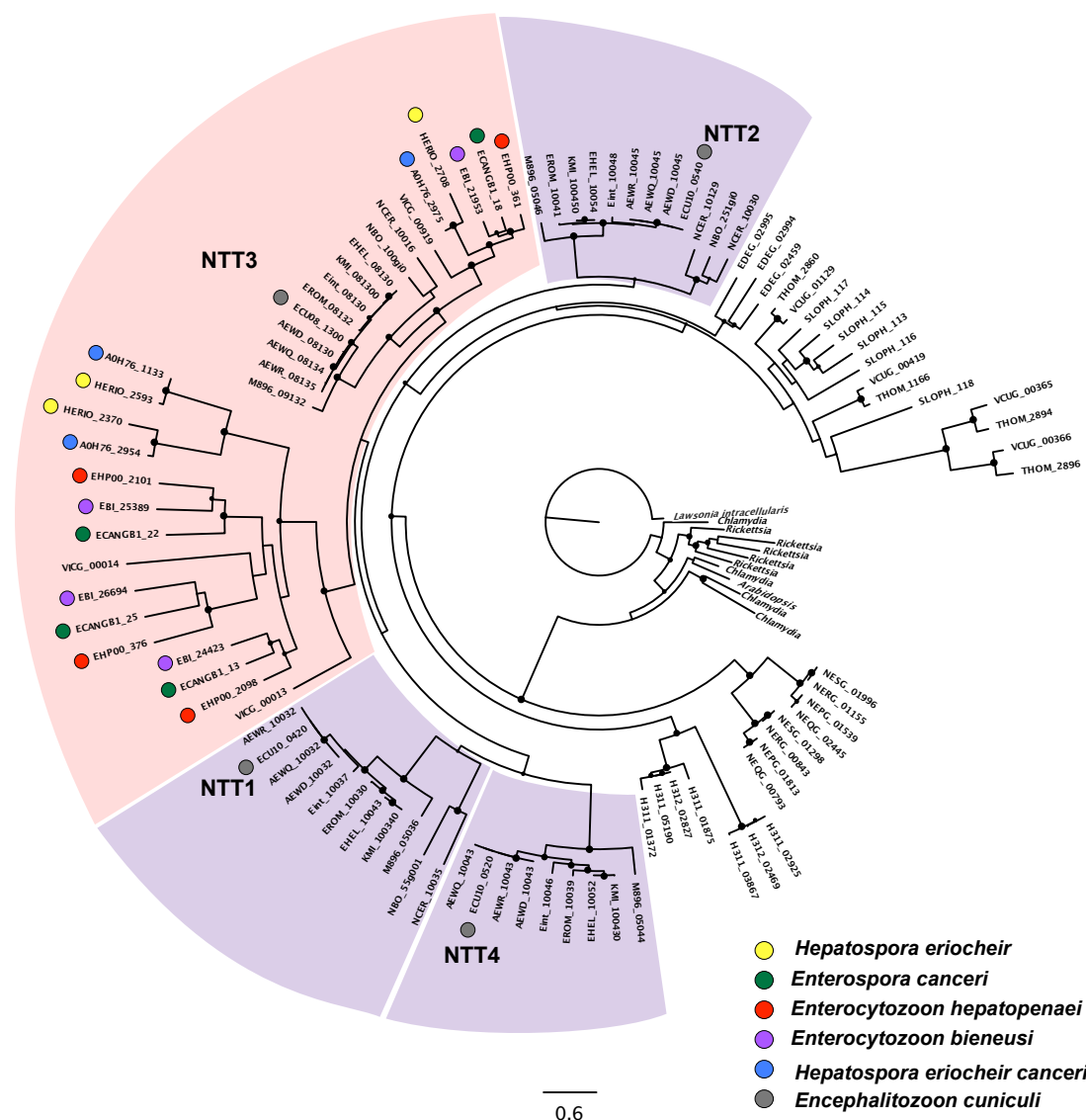


Figure 3.13: Maximum likelihood analysis performed on a masked alignment of microsporidian ATP/ADP translocases. Bootstrap inferences above 60 % have been represented by black dots on the respective branches. Prokaryotic homologs were used to root the tree. Protein clusters in which one of the four *Encephalitozoon cuniculi* homologs is present have been shaded. The pink shaded cluster represent the cluster containing ECU08_1300, the translocase that is thought not to be capable of NAD^+ transport (Tsaousis et al. 2008). No homolog for *Mitospordium daphniae* was identified in published genome.

3.4.7 Characterization of microsporidian hexokinases

3.4.7.1 Yeast functional complementation assay

The hexokinase triple knockout strain, YSH7.4-3C transformed with a GFP tagged chimeric hexokinase from *Ent. canceri* displayed several small colonies with a conventional round periphery which were confirmed by PCR to be positive transformants. A number of large colonies with a smudged periphery instead of the conventional smooth round periphery were also present on this plate. Following PCR analysis, it was evident that these larger colonies were contaminants. Further microsporidian analysis on transformed colonies showed two

types of GFP fluorescence: diffuse expression in the cytoplasm and punctate GFP expression. As expected no growth was observed in non-transformed YSH7.4-3C cells grown on glucose media. It was not possible to perform functional complementation assays due to inability of transformed YSH7.4-3C cells to grow in liquid YEPD/galactose media or when restreaked on agar plates.

3.4.7.2 MALDI-MS analysis of excised 50 and 25 kDa protein bands

The aim of this experiment was to confirm heterologous expression of the poly-histidine tagged (HIS-tagged) *Enc. cuniculi* hexokinase in *S. cerevisiae* model systems. The 52 kDa band (expected hexokinase size) sent for MALDI-MS analysis was identified as EF1 transcription factor protein and the intense 25 kDa protein band was identified as a ribosomal protein (Full MALDI-MS report in Appendix 8). This shows that the bands excised from the gel were not the expected hexokinase and this implies that the *Enc. cuniculi* recombinant protein was not the highly expressed protein band at 52 kDa. It is possible that the recombinant protein may not have been expressed in high volumes by the yeast cells and so it was difficult to have identified it from the protein gel. Considering the presence of a secretion signal at the N-terminus of the *Enc. cuniculi* hexokinase, it is likely that the recombinant protein may have been secreted into the growth media during overnight incubation. This was a pilot experiment to assess the efficiency of heterologous microsporidian hexokinase expression in a yeast model system. As the results here were not promising, an *Escherichia coli* model system was used instead.

3.4.7.3 Heterologous expression of microsporidian hexokinase in

Escherichia coli

In order to perform functional assays on the chimeric *Ent. canceri* hexokinase and to understand if the fused PTPA domain provides an added function, the protein, with and without its PTPA domain was heterologously expressed in *E. coli*. The hexokinase of *Enc. cuniculi* and of *H. eriocheir* were also included in this analysis. The chimeric *Ent. canceri* protein was in the range of 82 kDa whereas the same protein without its PTPA domain had a molecular weight of ~48 kDa. Hexokinase 2 of *H. eriocheir* was in the range of 50 kDa whereas the hexokinase of *Enc. cuniculi* was ~52 kDa. Among the four heterologously expressed proteins, that of *Enc. cuniculi* produced the faintest band (Figure 3.12). In summary, the heterologous hexokinase expression for all the above named species was

successful albeit at varying degrees. Considering the background noise on the Western blot (Figure 3.12) and the coomassie stained gel (not shown), it is evident that the extracted proteins were of low purity. In future experiments recombinant proteins purified by methods employed here could benefit from an extra dialysis step to improve sample purity.

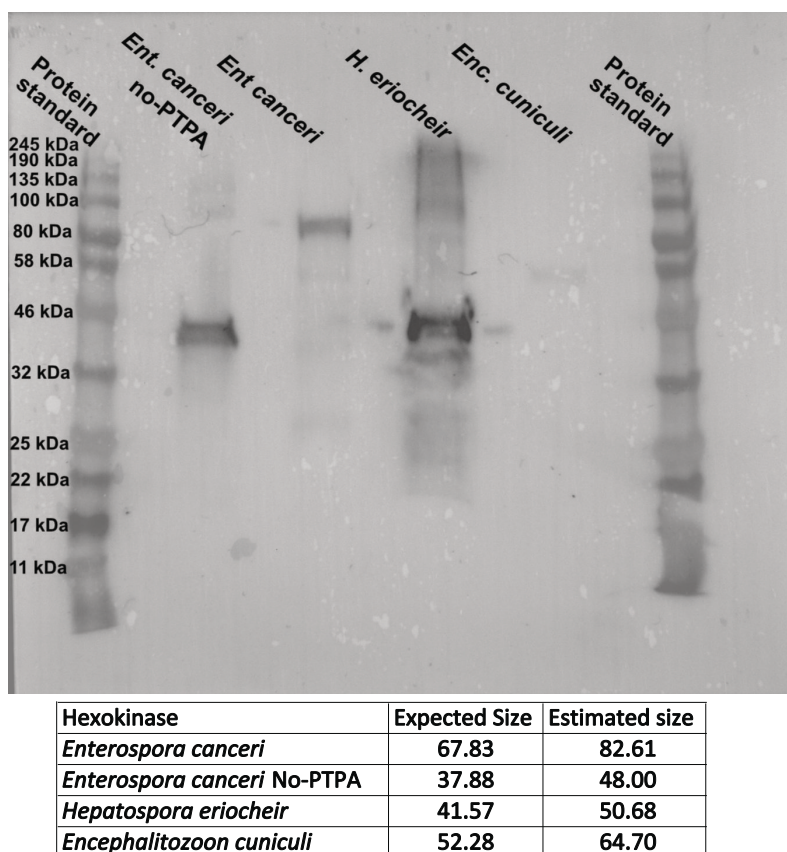


Figure 3.14: Western blot analysis of microsporidian hexokinases expressed in *Escherichia coli*. Hexokinases were detected by using anti poly-HIS mouse polyclonal antibody. The positions of molecular mass markers are indicated on both sides of the blot. Expected protein size and estimated protein size for each recombinant protein has been listed in the table below.

3.4.7.4 Hexokinase functional assay

To assess whether the recombinant microsporidian hexokinases could phosphorylate glucose, they were assayed in a coupled reaction containing glucose-6-phosphate dehydrogenase in the presence of NAD^+ and ATP. In this reaction, hexokinase phosphorylates glucose to glucose-6-phosphate by transferring a phosphoryl group from ATP onto a glucose molecule. The subsequent dehydrogenation of glucose-6-phosphate is catalysed by glucose-6-phosphate dehydrogenase, which uses NAD^+ as a co-enzyme. This coupled reaction releases NADH as a by-product, which is detectable at a wavelength of 340 nm. As the release of NADH is directly proportional to the amount of glucose-

6-phosphate dehydrogenase produced by hexokinase, NADH present in the reaction is a good measure of hexokinase's catalytic activity. It was not possible to heterologously express adequate amounts of *Enc. cuniculi* hexokinase for this experiment. Furthermore, no results were obtained for later experimental replicates in which *Ent. canceri*'s hexokinase without its PTPA were used. That is, in later experimental replicates, even the positive control (*S. cerevisiae*'s hexokinase 1) did not yield any results. Consequently, results from initial experiments performed only with recombinant hexokinases of *S. cerevisiae*, *Ent. canceri* (chimeric) and *H. eriocheir* are presented here. Absorbance measurement showed that the coupled reaction involving hexokinase from *H. eriocheir* hardly increased in NADH production with time. Although the coupled reaction involving the chimeric hexokinase of *Ent. canceri* produced high amounts of NADH that were comparable to that of the positive control (*S. cerevisiae* HK1), NADH increase was not recorded until 5 minutes after incubation. On the contrary, increase in NADH production could be recorded immediately after incubation for the positive control. NADH readings for the experimental samples were jagged unlike those for the positive control (Figure 3.15). It was not possible to test the kinetics of the recombinant enzymes in varying concentrations of glucose due to shortage of recombinant enzymes. Consequently the K_M and V_{max} of the enzymes could not be determined.

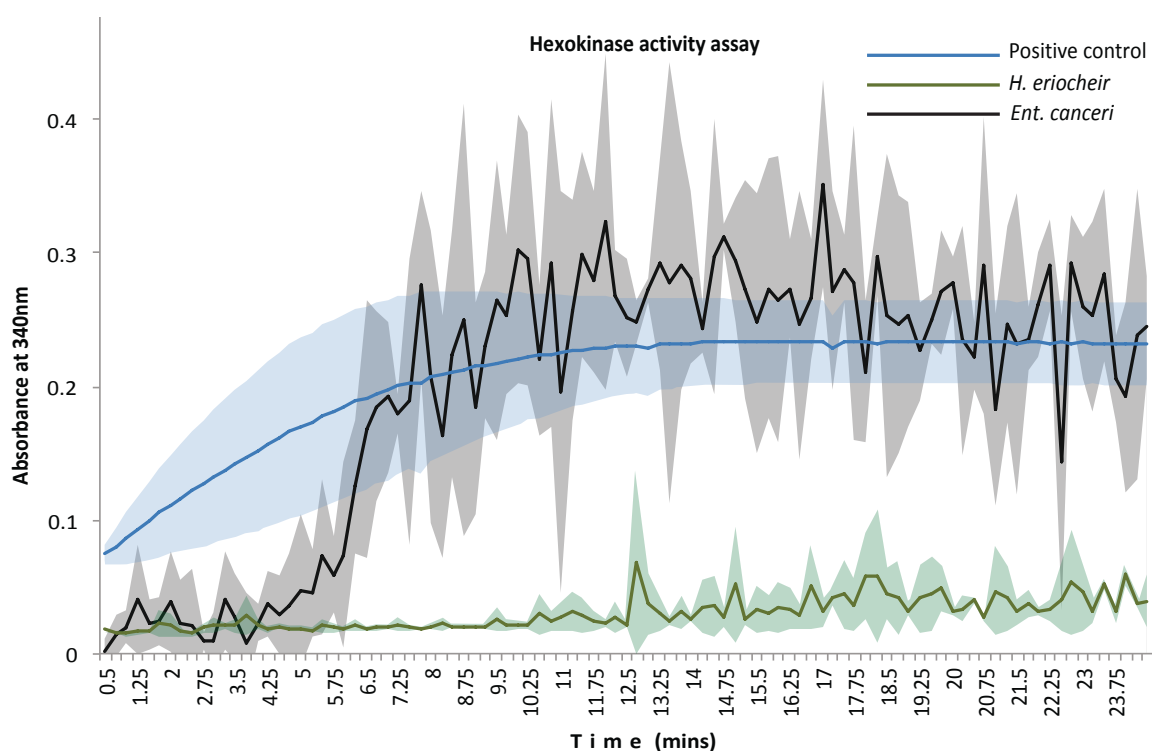


Figure 3.15: NADH absorbance readings for hexokinase/glucose-6-phosphate coupled reaction. Hexokinases are of *Enterospora canceri* and *Hepatospora eriocheir* heterologously expressed in *Escherichia coli*. Positive control is hexokinase 1 from *Saccharomyces cerevisiae*. Shadows represent standard deviation calculated from triplicates (*Ent. canceri*) and duplicate (*H. eriocheir* and *S. cerevisiae*) experiments.

3.4.7.5 BIOLOG phenotypic microarray

In an attempt to determine substrate specificity of the microsporidian hexokinases, *E. coli* cells transformed with hexokinases from *Ent. canceri*, *Enc. cuniculi* and *H. eriocheir* hexokinase 2 were grown on 95 different carbon sources and cell growth (NADH release) was monitored with a phenotype MicroArray system (BIOLOG, UK). On a glucose carbon source, *E. coli* transformed with *H. eriocheir* hexokinase had the longest lag growth phase with the log phase commencing 20 hours after inoculation. *E. coli* transformed with hexokinase from *Enc. cuniculi* and *Ent. canceri* (chimeric) displayed similar lag phases although cells transformed with *Enc. cuniculi*'s hexokinase entered a second transient lag phase a few hours into their log phase. After 60 hours, *E. coli* transformed with *Ent. canceri*'s chimeric hexokinase displayed more growth than the other two experimental and control samples (*E. coli* transformed with an empty vector) (See pink box in Figure 3.16). Similar results were observed for cells growing in other hexose sugars such as mannose and fructose (See purple box in Figure 3.16). Here although the log growth phase for *E. coli* transformed with *Ent. canceri*'s chimeric hexokinase repeatedly commenced earlier than cells transformed with hexokinases from *Enc. cuniculi* and *H. eriocheir*, there was minimal difference in growth observed between all three experimental and control samples after 60 hours. In a number of other carbon compounds such as citric acid, galactose, myo-inositol, glucosamic acid, glutamine and propionic acid, the log phase of bacteria transformed with *Ent. canceri*'s chimeric hexokinase commenced earlier than the other two experimental samples and positive control (Figure 3.16).

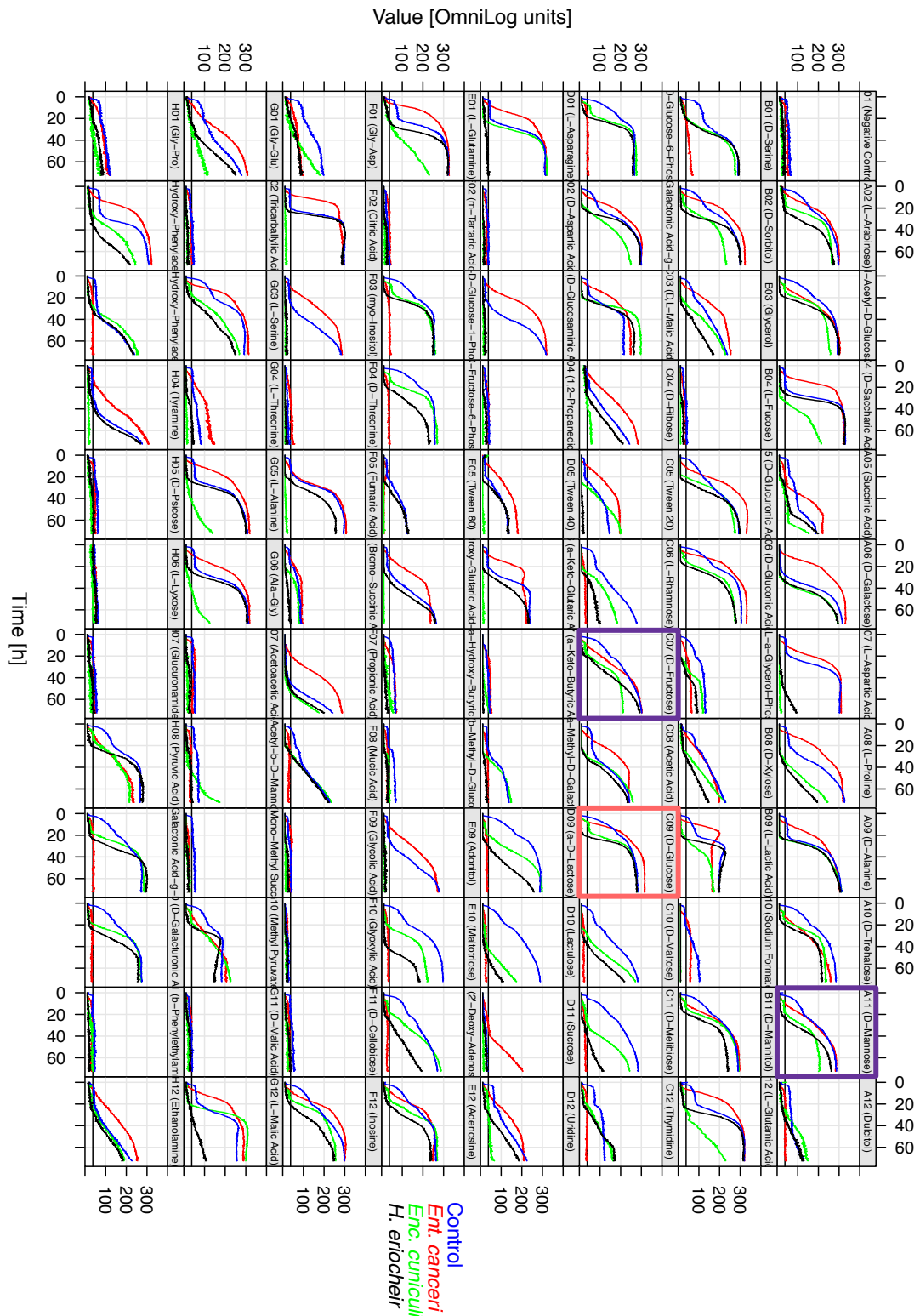


Figure 3.16: BIOLÓG Phenotypic microarray analysis of Rosetta(DE3)pLysS *E. coli* cells transformed with hexokinases from three different microsporidian species: *Ent. canceri* (full hexokinase protein including PTPA domain), *Enc. cuniculi* and *H. eriocheir*. Cells were grown on a PM-1 microwell plate containing 95 different carbon sources including glucose (pink box), mannose and fructose (purple box).

3.5 Discussion

3.5.1 Absence of glycolysis within the Enterocytozoonidae

The genome survey of the human pathogen *E. bienersi* by Nakjang *et al.* (2009) suggested that this microsporidian lacked a glycolytic pathway. The peculiarity of this discovery warranted a second sequencing attempt the following year, which yielded similar results (Keeling *et al.* 2010). Confirmation of the absence of a complete glycolytic pathway in *E. bienersi* in turn sparked questions of whether this is a common phenomenon in the Microsporidia (Keeling *et al.* 2010). Comparative whole genome analysis performed here corroborate Keeling's findings but also demonstrate that species closely related to *E. bienersi* have lost glycolytic capabilities (Figure 3.3). That is, genomes of all members of the Enterocytozoonidae family sequenced in this study, namely *Ent. canceri*, *E. hepatopenaei*, *H. eriocheir* and *H. eriocheir canceri* lacked the full set of genes needed for a complete glycolytic pathway.

3.5.2 Absence of glycolytic genes in sequenced genomes is not due to incomplete genome sequencing

Although none of the genomes presented in this study were sequenced to completion, it is unlikely that missing glycolytic genes are as a result of incomplete sequencing of the genome. Out of 44 conserved microsporidian transcriptional control proteins, genomes of both *Hepatospora* species encoded for 40 proteins whereas those of *Ent. canceri* and *E. hepatopenaei* encoded for 42 and 43 respectively. Using these figures as a measure of completeness, genomes sequenced in this study can be considered to be 84 to 100 % complete and the probabilities of observing this level of missing glycolytic genes simply as a result of incomplete sequencing (rather than true absences) is 2.869×10^{-08} for *Ent. canceri*, 6.975×10^{-10} for *E. hepatopenaei*, 8.629×10^{-05} for *H. eriocheir canceri* and 8.629×10^{-05} for *H. eriocheir*. In addition to this statistical support, there are a number of factors that corroborate the assumption that the absence of glycolytic genes observed in the above-mentioned genomes are not a consequence of incomplete genome sequencing. Firstly, the genome of *E. hepatopenaei* was independently sequenced using PACBIO and MISEQ technologies, and assemblies from both sequencing attempts lacked a complete set of glycolytic genes. Similarly, the same pattern of glycolytic gene loss was observed for the genomes of the two *Hepatospora* sub species even though they

were isolated from different hosts, sequenced with different protocols and assembled independently.

3.5.3 Explaining events that led to the decay of glycolysis in the Enterocytozoonidae

The differential loss of glycolytic genes displayed by the Enterocytozoonidae genomes analysed here suggests the selective pressures that maintain this pathway became relaxed at the base of this clade and resulted in the parallel decay of the pathway in extant species (Figure 3.3). In the Microsporidia, glycolysis has been postulated to be the main ATP source for the spore stage, especially during germination (Dolgikh et al. 2011; Heinz et al. 2012). This makes the absence of glycolysis within member species of the Enterocytozoonidae family presented in this study intriguing. Below, data produced in this study is dissected and compared to published work to attempt at an explanation of the events that culminated in the loss of this pathway in the Enterocytozoonidae.

3.5.3.1 Horizontal transfer of ATP/ADP translocases in the Microsporidia

The Microsporidia have horizontally acquired ATP/ADP translocases from Chlamydia-like ancestors that are capable of transporting ATP (Richards et al. 2003; Tsaousis et al. 2008; Heinz et al. 2014) (Figure 3.13). As expected, all Enterocytozoonidae genomes sequenced in this study encoded for 3 or 4 copies of this protein, hinting to the indispensability of these transporters in these metabolically reduced lineages (Figure 3.2) (Appendix 10). The horizontal acquisition of these ATP/ADP translocases likely happened in the last common ancestor of the Cryptomycota and the Microsporidia (Richards et al. 2003; Tsaousis et al. 2008; Heinz et al. 2014) (Figure 3.13). As such, if host ATP import by these ATP/ADP translocases alone was enough to relax the selective pressure to maintain glycolysis, one would have envisaged all microsporidians genomes sequenced till date to display a similar loss of glycolytic genes as displayed by Enterocytozoonidae genomes but this is not the case. That is, although the horizontal acquisition of ATP/ADP translocases may have partly contributed to the eventual decay of the glycolytic pathway, it is unlikely that it was the only contributing factor.

3.5.3.2 Loss of cytosolic NAD⁺ pool replenishing pathways

3.5.3.2.1 AOX enzymes

In 2010, Williams *et al.* published data demonstrating that genomes of some microsporidian lineages encoded AOX proteins. The authors suggested that AOX, in conjunction with glycerol-3-phosphate dehydrogenase may be responsible for the conversion of NADH, a bi-product of glycolysis into NAD⁺ and feeding it back into the glycolytic pathway. Thus, these enzymes form part of a metabolic cycle that replenish NAD⁺ pools in the microsporidian cell and ultimately permit the sustainability of glycolysis (Williams et al. 2010) (see Figure 3.1 for role of NAD⁺ in glycolysis). Comparative genomic data presented in this study shows that although glycerol-3-phosphate dehydrogenase is present in all analysed genomes, AOX is absent in all aglycolytic lineages and in a subset of lineages in which glycolysis is present, namely *Vittaforma*, *Anncaliia*, *Nosema*, *Ordospora* and *Encephalitozoon* genera (Figure 3.4). Horizontal gain of ATP/ADP translocases in addition to the absence of a metabolic machinery to maintain cytosolic NAD⁺ pools may have indeed relaxed selective pressure to maintain glycolysis in the Enterocytozoonidae. However, the absence of AOX enzymes in a subset of lineages that have retained their glycolytic pathways suggests that the relaxation of selective pressure to maintain glycolysis at the base of the Enterocytozoonidae tree is as a consequence of more than just horizontally acquired ATP/ADP translocases and absence of AOX enzymes.

3.5.3.2.2 Loss of the mevalonate biosynthesis pathway

There must have been the loss or gain of a gene/genes at the base of the Enterocytozoonidae tree that led to the relaxation of selective pressure to maintain glycolysis in the Enterocytozoonidae. However, I hypothesise that this event is likely to be a loss of gene/genes rather than a gain due to the general reductive trajectory of microsporidian genome evolution (Corradi et al. 2010; Keeling et al. 2010; Nakjang et al. 2013). To test this hypothesis, orthologous protein clusters that were exclusive to glycolytic microsporidians but absent in aglycolytic species were identified. As expected some of these protein clusters were glycolytic proteins that were absent in the Enterocytozoonidae. However a number of these protein clusters were protein families of enzymes involved in the mevalonate biosynthesis pathway (Figure 3.5). This pathway is important for synthesizing isopentenyl pyrophosphate (IPP in Figure 3.5), which is used for

protein prenylation, cell membrane maintenance, terpenoid synthesis, protein anchoring, and N-glycosylation (Russell 1992; Berg et al. 2006; Jones et al. 2009). With regards to glycolysis, the mevalonate biosynthesis pathway produces a net NAD^+ output of 1 per pyruvate molecule. Since glycolysis produces 2 pyruvate molecules per glucose molecule, it implies that for every glucose molecule metabolised in glycolysis, the mevalonate synthesis pathway releases 2 free NAD^+ molecules. In theory, this is enough to tally glycolytic NAD^+ requirement of 2 NAD^+ per glucose molecule (Figure 3.5). Considering its likely role in maintaining cytosolic NADH/NAD^+ balance, the loss of this pathway prior to the radiation of the *Vittiforma/Enterocytozoonidae* clade may have contributed to the reduction of selective pressure to maintain glycolysis in the *Enterocytozoonidae* (Figure 3.4).

3.5.3.2.3 Loss of NUDIX enzymes

Apart from proteins involved in the mevalonate pathway, bis(5'-nucleosyl)-tetraphosphatase was among the protein clusters that were absent in the *Enterocytozoonidae* but present in glycolytic lineages. Bis(5'-nucleosyl)-tetraphosphatases belong to the NUDIX (nucleoside diphosphate linked to some moiety, X) protein family. Members of this protein family have been implicated in a plethora of cellular functions including, the regulation of ATP-sensitive channels, cell invasion and activation of gene expression (Jovanovic et al. 1997; Martín et al. 1998; Ismail et al. 2003; Lee et al. 2004). Recently, homologs of this protein family in *Arabidopsis thaliana* have been implicated in the recycling of NADH to NAD^+ (Ishikawa et al. 2009; Ogawa et al. 2016). If this is also true for microsporidian homologs, then they may contribute to the maintenance of homeostatic balance between NAD^+ and NADH . This explanation is particularly plausible as NUDIX genes were present in almost all glycolytic species lacking AOX enzymes (Figure 3.4). Again, this gene was lost at the base of the *Vittiforma/Enterocytozoonidae* clade, hinting to its likely contribution to the relaxation of selective pressures to maintain glycolysis. Apart from the above-mentioned protein families, the remaining orthologous clusters identified in this study that were exclusive to glycolytic lineages could not be assigned a function by BLAST-based methods. It is possible that the absence of the mevalonate synthesis pathway and NUDIX enzymes in the *Enterocytozoonidae* may have not played a role in the eventual erosion of glycolysis and may in fact just be as a

consequence of other evolutionary factors. However, the indispensability of NAD⁺ for the glycolytic pathway and the likely involvement of the mevalonate synthesis pathway and NUDIX enzymes in the NAD⁺/NADH homeostatic balance warrants further research to be directed here.

3.5.3.2.4 Loss of NAD⁺ transport systems

As mentioned earlier, microsporidian ATP/ADP translocases were horizontally acquired from a *Chlamydia*-like endosymbiont in the last common ancestor between the Cryptomycota and the Microsporidia (Richards et al. 2003; Tsaousis et al. 2008; James et al. 2013; Heinz et al. 2014) (Figure 3.13). There has however been recent reports that ATP/ADP translocases in some extant chlamydiales are able to transport NAD⁺ with high specificity as well as ATP (Haferkamp et al. 2004; Fisher et al. 2013). NAD⁺ has also been demonstrated to be a competitor of ATP for three of the four ATP/ADP translocases of *Enc. cuniculi* when expressed in *E. coli* (Tsaousis et al. 2008). Interestingly, in phylogenetic analysis performed here ATP/ADP translocases belonging to members of the Enterocytozoonidae clustered specifically with the *Enc. cuniculi* homolog that did not recognise NAD⁺ as a substrate, ECU08_1300 (Figure 3.13). Although this needs to be corroborated by experimental data, the clustering of Enterocytozoonidae ATP/ADP translocases with only NAD⁺-non-sensitive and not with NAD⁺ sensitive *Enc. cuniculi* homologs suggests that the Enterocytozoonidae homologs may be insensitive to NAD⁺ and hence cannot import this substrate from their hosts. Absence of such a system to replenish cytoplasmic NAD⁺ pools in the Enterocytozoonidae may have made glycolysis unsustainable in this lineage and led to a reduced selective pressure to maintain this pathway. This perhaps culminated in the decay of glycolysis during speciation of members of the Enterocytozoonidae.

3.5.3.3 Summary of events that may have led to loss of glycolysis in the Enterocytozoonidae

The above-discussed factors that may have led to the loss of glycolysis in the Enterocytozoonidae have been summarized below:

- Loss oxidation phosphorylative capability at base of microsporidian phylum

- Lateral gain of ATP/ADP translocases in the ancestor of the Microsporidia and Cryptomycota.
- Increased capacity of ATP import from host via ATP/ADP translocases caused a decrease in demand for intrinsic ATP production by glycolysis.
- This culminated into a reduced selective pressure to maintain genes involved in glycolysis especially “backend” genes such as AOX.
- Absence of backend genes made glycolysis unsustainable.

Eventually, relaxed selective pressure to keep glycolytic genes led to the loss of these genes in the Enterocytozoonidae with the exception of hexokinase due to its key role in other metabolic pathways. Data presented here suggests that not all glycolytic enzymes may have been lost via species-specific erosion of the pathway. There is the possibility that four (phosphofructokinase, fructose-bisphosphate aldolase, phosphoglycerate kinase and pyruvate kinase) out of the nine glycolytic enzymes that are absent in the Enterocytozoonidae were lost at the base of the family during a single event (Figure 3.3). This can be verified in the future when more genomes of member species of the Enterocytozoonidae family become available.

3.5.4 Microsporidian hexokinases

The retention of the hexokinase gene in all microsporidians, even in aglycolytic lineages (Figure 3.3) hints to the presence of a selective pressure to maintain it and even duplicate it in some species (Figure 3.5). This is potentially due to its dynamic role in metabolism (Figure 3.2). That is, hexokinase is not only involved in glycolysis but also in the pentose phosphate pathway and mannose metabolism (Figure 3.2) (Berg et al. 2006). In phylogenetic analysis performed here, hexokinases of microsporidians branched together with the *Rozella* homolog and the glucokinase of *S. cerevisiae* rather than branching at the base of the *S. cerevisiae* hexokinase/glucokinase clade (Figure 3.6) (Figure 3.7). This is interesting as it hints to the presence of a single glucokinase (with no hexokinases) in the primordial ancestor shared between canonical Fungi, Rozellids and the Microsporidia. Thus diversification of hexokinase 1 and 2 in canonical Fungi occurred after the radiation of the microsporidian and Rozellid lineages. This is in agreement with published phylogenetic and structural studies performed on eukaryotic hexokinases (Bork et al. 1993; Cornish-bowden et al. 1998). Results presented in this study demonstrate that the primordial

microsporidian genome encoded a single hexokinase, which may have been a glucokinase (Figure 3.6 and 3.7), and that the multiple hexokinases encoded by extant microsporidian genomes are as a result of recent, lineage specific duplication events (Figure 3.6 and 3.7). Gene duplications of kinases in eukaryotic genomes are common (Grossbard & Schimke 1966; Ramel et al. 1971; Suga et al. 1999; Blin et al. 1999; Conant & Wagner 2002) and gene duplication in general is known to facilitate innovation in genomes by allowing the duplicate gene to develop new functional properties via the accrual of non-deleterious mutations, a process referred to as neofunctionalization (Force et al. 1999; Lynch et al. 2001; van Hoof 2005). In cases, where the new gene copy accrues deleterious mutations, it is either pseudogenized (rendered non-functional) or lost (Stoltzfus 1999). It has also been postulated that duplications could lead to functional specialization of the duplicated genes as observed for the two hexokinase in yeast (James & Tawfik 2003; Oakley et al. 2006). Another important role gene duplication plays in genome evolution is the provision of extra templates for the bulk production of important proteins. This is referred to in literature as the gene dosage model and has been suggested to best fit genes involved in metabolism such as hexokinase (Brown et al. 1998; Guillemaud et al. 1999; Kondrashov et al. 2002). Figure 3.6 shows that although *E. bienewisi* have multiple copies of the hexokinase gene, all but hexokinase 1 have accrued mutations at their active sites. This alludes to the retention of the ancestral function by hexokinase 1 and the possible pseudogenization of the duplicate genes. Similarly, hexokinase 2 in *Nos. bombycis*, and hexokinase 3 and 4 in *T. hominis* are likely to be pseudogenes due to heavy mutations at their active sites (Figure 3.6). As such, it is possible that for these genomes, duplicate hexokinases may be non-functional or may have acquired new functions. The type of hexokinase duplication observed for the *Hepatospora* spp. perhaps presents a Gene Dosage Model as both gene copies present similar mutations.

In line with findings by Cuomo *et al.* (2012), analysis performed here demonstrate that a number of microsporidian hexokinases possess an N-terminal peptide sequence that targets this protein to the secretory pathway (signal peptide) (Figure 3.6). Presence of a signal peptide appeared to be very species-specific. For instance *Enc. cuniculi* GB possessed a signal peptide whereas other sub-strains of the same species did not (Figure 3.6). In their study, Cuomo *et al.*, (2012) hypothesised that hexokinase targeted to the secretory pathway in the

Microsporidia could be secreted by meronts to boost ATP production in the host which can be then pilfered by the parasite via its ATP/ADP translocases. Hexokinase gene copies of *T. hominis* and *Vav. culicis* also possessed a signal peptide. Thus these two species, have specialised copies of their hexokinase gene to exploit the host environment while retaining a copy of the gene for intrinsic metabolism.

Hexokinases belonging to all the Enterocytozoonidae family members presented in this study (including *E. bienewisi*) did not possess signal peptides. Instead, the N-terminal of the hexokinase of *Ent. canceri* appeared to be fused to a protein phosphotyrosyl phosphate activator (PTPA) domain. A scan of proteins encoded by all the microsporidian genomes analysed in this study revealed that formation of chimeric proteins is uncommon in extant members of this phylum but may have been abundant in the common ancestor of the Microsporidia (Figure 3.12). Considering the unusual occurrence of chimeric proteins in the microsporidia, especially in lineages that radiated more recently such as the Encephalitozoonidae and the Enterocytozoonidae (Figure 3.12), a gene order conservation survey was performed to assess the provenance of both the hexokinase and PTPA domains of the *Ent. canceri* chimera. This analysis demonstrated that both domains existed as distinct genes, and were probably situated on separate chromosomes in the common ancestor of the Enterocytozoonidae. Insertion of the hexokinase gene in close proximity to the PTPA gene appears to be a recent event that ensued at the base of the *Enteocytozoon/Enterospora* clade. This insertion may have fostered the formation of the hexokinase chimera observed in *Ent. canceri* (Figure 3.9-11). This process of gene juxtaposition followed by fusion seems to be a common mechanism adapted by genomes to introduce functional innovation as it has been observed in hominoids, fruit flies, *Plasmodium* and trypanosomes (Cortés 2005; Zhou et al. 2008; Bartholomeu et al. 2009; Marques-Bonet et al. 2009). The significance of the formation of the PTPA-hexokinase chimera in *Ent. canceri* is unclear but the conservation of all active sites in the hexokinase domain suggests this protein may have retained its ancestral hexose-phosphorylating function. The potential role of the fused PTPA domain may be to regulate the function of hexokinase. Typically, the PTPA domain would have the capacity to remove a phosphate from a phosphorylated tyrosine. In other organisms, proteins containing this domain are involved in the regulation of the function of other

proteins (Janssens & Goris 2001). This may suggest that this chimera now performs a dual role as both a protein phosphatase and a hexokinase. Further experiments need to be conducted to ascertain the function of this enigmatic chimera.

Among the 50 hexokinases analyzed in this study, those belonging to *Hepatospora* species were the only homologs that had undergone an E302H amino acid substitution. This was peculiar because histidine (H) has a basic side chain whereas glutamic acid (E) has an acidic side chain. All substitutions at this site in other microsporidian hexokinases were of amino acids that possessed acidic side chains, Aspartic acid (D). Active site substitutions with amino acids with similar side chains often do not alter protein function whereas substitutions with amino acids with dissimilar side chains do (Cupples & Miller 1988). As this mutation occurred in the glucose-binding domain, one is inclined to hypothesize that hexokinase homologs of *Hepatospora* may not recognize glucose as a substrate as conventional hexokinases do. This hypothesis is supported by preliminary functional characterization data discussed below.

3.5.5 Functional characterization of hexokinase from *Enterospora canceri* and *Hepatospora eriocheir*

In order to assess the specificity of microsporidian hexokinases for different carbon sources, recombinant hexokinases expressed in *E. coli* cells were purified and a functional assay was performed on them. Yield for *Enc. cuniculi*'s hexokinase was low and so it was not possible to perform functional assays on it. Low yield for *Enc. cuniculi* was understandable as bioinformatic predictions hints towards a signal peptide at the N-terminus of this protein suggesting that this protein may be secreted out of the cell. There is experimental data to suggest that hexokinases possessing signal peptides in other microsporidian species are secreted into the host cell (Senderskiy et al. 2014). Preliminary functional assays performed on recombinant hexokinases of *Ent. canceri* and *H. eriocheir* revealed that the chimeric homolog of *Ent. canceri* may recognise glucose as a substrate whereas that of *H. eriocheir* may not. The 5 minute delay prior to an observable rise in catalytic activity by *Ent. canceri*'s chimeric hexokinase may have been due to temperature sensitivity of the protein (Figure 3.15). Thus this enzymatic assay was designed in a way that addition of the recombinant enzyme (kept on ice) to the enzymatic reaction was the final step prior to the incubation of the assay plate

in a microplate reader. Consequently, the 5 minute delay may have been time needed for the assay's temperature to reach room temperature. Observation of this time delay in only the assay performed with *Ent. canceri*'s chimeric hexokinase and not in other hexokinases suggests this is a chemical property of the enzyme itself. It is likely that the gentle shaking of the assay plates prior to each reading may have introduced bubbles into the samples causing the erratic readings observed in Figure 3.15. It is however difficult to envisage how this would only affect the experimental sample and not the positive control since all samples were assayed on the same plate. Another plausible explanation for these erratic readings is that the salts present in the buffer in which the recombinant proteins were eluted may have interfered with the absorbance readings. In future experiments, dialysis of eluted recombinant proteins with buffers that contain less salt may be required to circumvent this problem. It was not possible to perform further replicates of this experiment to determine the enzyme kinetics of the microsporidian hexokinases due to shortage of recombinant proteins and unsuccessful heterologous expression and purification of additional recombinant proteins.

The rationale behind using BIOLOG assays in this study was to monitor how *E. coli* cells heterologously expressing microsporidian hexokinases grow on different carbon sources. Thus providing a high throughput system to assess microsporidian hexokinase substrate specificity. Faster growth observed in the control strain for all six-carbon sugars as compared to experimental strains was odd although not entirely unexpected (Figure 3.16). This is because although the control strain was transformed with an empty plasmid and so did not contain a microsporidian hexokinase, it did possess its own bacterial hexokinase gene, which would have enabled it to grow on six carbon sugars. Experimental strains may have been burdened with the extra task of expressing the recombinant microsporidian hexokinases thereby delaying their growth (Figure 3.16). Growth was often not detected for *E. coli* transformed with hexokinases from *Enc. cuniculi* and *H. eriocheir* until 20 hours after incubation. As the stability of ampicillin, the selective marker used in this study is dramatically reduced when incubated for this amount of time at the temperature range used in this assay (Zhang & Trissel 2002), it is likely that most of the growth observed after the 20 hour mark is as a result of contamination.

It was interesting to observe that experimental strains transformed with the chimeric hexokinase from *Ent. canceri* displayed faster growth on some carbon sources as compared to strains transformed with other microsporidian hexokinases (Figure 3.15). These carbon sources included hexose sugars (D-Glucose, D-Galactose, D-Fructose, D-Mannose), oxidised hexose sugars (D-Gluconic acid, D-Saccharic acid), pentose sugars (D-Xylose, Thymidine), weak acids (L-Lactic acid, Citric acid) to amino acids (Glutamic acid, Proline, Serine). Due to the probable noise introduced by metabolism performed by the bacteria's own hexokinases, this experiment was abandoned and replicates were not performed. It is therefore not possible to conclude whether the elevated growth observed for cells transformed with *Ent. canceri*'s hexokinase was as a true consequence of the recombinant protein or of contamination in the sample. In previous studies, contaminated samples have been recorded to produce similar anomalous results with rapid growth attributed to contamination (Khatri et al. 2013).

In order to circumvent the problem of noise introduction by native hexokinases from the model organism being used, a triple hexokinase *S. cerevisiae* mutant strain was used. That is, to determine whether the recombinant microsporidian hexokinases could recognise glucose as a substrate, they were tested to see if they could restore growth in an *S. cerevisiae* triple hexokinase knock-out mutant strain growing on a glucose substrate. Since these strains lack endogenous hexokinase activity, they do not grow on selection medium containing glucose as the sole carbon source (Hohmann et al. 1999). Hence, complementation assays such as the one performed in this study can be used to assess the functionality of recombinant hexokinases (Cho et al. 2006; Nilsson et al. 2011). Untransformed strains were difficult to grow on galactose, the recommended carbon source by the authors (Hohmann et al. 1999). Following transformation with the *Ent. canceri* chimeric hexokinase, growth on glucose was observed however, colonies were smaller than expected and repeats of this experiment produced inconsistent results. Furthermore, no growth was observed for positive transformants that were restreaked or inoculated in liquid media. This was strange as PCR screening had confirmed these colonies as positive transformants. Microscopy analysis showed that some cells expressed the GFP tagged protein albeit weakly. Again, a GFP signal was observed only for a minority of the cells analysed. Overall these results are inconclusive and further repeats of the experiments are

needed to verify the functionality of the protein when expressed in *S. cerevisiae*. Past studies have demonstrated that the GFP tag could alter the biology of the tagged recombinant protein (German-Retana et al. 2000; Skube et al. 2010)(Mahen et al. 2014). Future experiments could benefit from the use of a smaller affinity tag such as a poly-histidine tag rather than a GFP tag used here. In such studies, the recently described Ni-NTA-AC fluorescent probe, which targets HIS-tagged proteins in live cells could be used to observe the subcellular localization of the recombinant hexokinases (Lai et al. 2015). Functional complementation experiments for hexokinases from *H. eriocheir* and *Enc. cuniculi* still remain to be performed. In the future, it will also be of use to assess whether the hexokinase of *Ent. canceri* without its PTPA domain is also able to complement triple knock-out yeast cell growth on glucose media.

3.6 Conclusion

As a whole the data provide strong evidence of the consistent loss of glycolysis into the Enterocytozoonidae. This data also shows that glycolytic enzymes were not lost in a single event in the ancestor of the group but rather there has been a common loss of the selective pressure to retain glycolysis followed by a parallel erosion of the pathway by differential loss of the enzymes across the members of the genus. Microsporidia are already primary models of metabolic reduction in eukaryotes, the Enterocytozoonidae take this reduction even further making them the only eukaryotic group to have eliminated all canonical ATP-generating pathways. Some of the analysis performed in this chapter have been submitted for publication to PLoS Pathology (See Wiredu-Boakye *et al.* 2016 in appendix 13).

Chapter 4 *Hepatospora*-An example of plasticity in microsporidian morphology and karyotype

4.1 Introduction

Since the inception of the phylum “Microsporidia”, the classification of this group has proven problematic at both higher and lower taxonomic levels [reviewed in (Corradi & Keeling 2009)](Section 1.3). The classification conundrum observed at higher taxonomic levels was due to the apparent amitochondrial nature of microsporidians and the position of this phylum as early branching eukaryotes on distance-matrix (Vossbrinck et al. 1987) and maximum likelihood phylogenetic trees (Kamaishi, Hashimoto, Nakamura, Nakamura, et al. 1996). Now, evidence has however shown the presence of a relic mitochondria in microsporidian cells (Williams et al. 2002) and that the positioning of microsporidia as early branching eukaryotes (Hirt et al. 1999) in early phylogenetic trees was artefactual (Vossbrinck et al. 1987; Kamaishi, Hashimoto, Nakamura, Nakamura, et al. 1996).

The latter was as a result of the inability of the phylogenetic tools used by previous authors to account for rate heterogeneity among gene sites, base-compositional biases and the overall accelerated evolutionary rate characteristic of microsporidian genomes (Hirt et al. 1999). Recent studies using phylogenetic probabilistic models that account for the above mentioned genome peculiarities (James et al. 2006) and synteny studies between canonical fungi and microsporidia (Lee et al. 2008) have led to the current placement of this phylum as highly derived fungi. Presently, there is some evidence hinting that the Microsporidia may have a phylogenetic affiliation with the newly created Cryptomycota group situated at the base of the fungal tree of life (Keeling 2014). As well as placing the microsporidian phylum in the tree of life, there have been problems in the classification of taxa within the actual microsporidian phylum. Classical systematics within this phylum as with many other phyla at the time was based on ultrastructural morphological features (Vávra & Undeen 1970; Cali et al. 1993; Shadduck et al. 1990; Canning 1953). This reliance on morphological features for the internal classification of the Microsporidia was due to the lack of computational resources and molecular data currently available. Within the past 25 years however, taxonomy has moved towards an integrative approach of using both morphological and molecular evidence. This has seen the revision of

many phyla and even led to the discovery of cryptic species. The study by Vossbrinck *et al.* (2005) is the most recent of its kind to use rDNA molecular evidence to revise the entire microsporidian phylum. This molecular marker has been invaluable in overhauling the microsporidian phylogeny to its current state and informing taxonomic nomenclature.

Due to the usefulness of rDNA in providing data for phylogenetic studies, their conserved nature has been harnessed to design a set of universal primers specific to this phylum (f18 and 1492R) (Vossbrinck *et al.* 1993; Kent *et al.* 1996). There are currently 3,184 rDNA microsporidian sequences on the NCBI database (Tatusova *et al.* 2014) as a result of three decades of using microsporidia-specific rDNA primers. This database provides a rich resource for assessing the general taxonomic position of novel members of this phylum (Vossbrinck & Debrunner-Vossbrinck 2005; Winters & Faisal 2014).

Nonetheless, microsporidian rDNA sequences obtained by using the above-mentioned universal primers are only able to resolve taxonomic relationships only up to the genus level (Vossbrinck & Debrunner-Vossbrinck 2005; Vossbrinck *et al.* 1998). Thus, the maximum nucleotide number of approximately 1400 bp amplified by these primers does not provide enough data for phylogenetic tools to distinguish between closely related species. Some authors were able to design primers that could target the variable ITS region however these were only specific to members of the Encephalitozoonidae family (Vossbrinck & Debrunner-Vossbrinck 2005; Vossbrinck *et al.* 1993).

An important feature of rDNA is its tandem repetitive arrangement in eukaryotic genomes [reviewed in (Dover & Coen 1981)]. This characteristic may have compromised the utility of rDNA in phylogenetic analysis if it was not for its sequence homogenisation via concerted evolution. This phenomenon gives rise to higher sequence similarity between repeats within genomes than between genomes (Dover & Coen 1981). Therefore, sequences obtained by primer amplification of a tandem from a genome would still be unique to that genome regardless of the number of repeats present. It has however come to light that some microsporidian species exhibit a dispersed arrangement of their rDNA units across multiple chromosomes as compared to conventional tandem arrangement (Liu *et al.* 2008). For this reason, the sequence homogenisation between repeats owing to concerted evolution is reduced leading to nucleotide variations within species and rDNA subunit rearrangement between members of the same genus

(Ironsides 2013). In his findings, Ironsides (2013) observed a higher ITS sequence variation between repeats in the same genome than between different genomes for some *Nosema* species. He goes on to question whether ITS regions should indeed be used as universal fungal markers as proposed by some authors (Schoch et al. 2012) and joins Pombert *et al.* (2013) to suggest the use of protein coding genes in the assessment of closely related species (Pombert et al. 2013). The use of protein-coding genes have been pivotal in the reassignment of the microsporidian phylum as a fungal group in the tree of life (Hirt et al. 1999; Keeling 2003; Fast et al. 1999; Vivarès et al. 2002) but only two authors so far have used this method to distinguish between closely related microsporidia (Haag et al. 2013; Brown et al. 2010). A comparison of the effectiveness between protein-coding and rDNA data in resolving close taxonomic relationships in fungi have demonstrated the superiority of the former over the latter (Hofstetter et al. 2007). More importantly, this study recommended the use of multi-loci phylogenetics for the resolution of phylogenetic relationships in fungi (Hofstetter et al. 2007). The problem with previous multi-loci microsporidian phylogenies was their inability to take distinct rates of individual gene evolution into account which consequently led to systematic errors in phylogenetic estimations (Keeling et al. 2005). Another disadvantage of using multiple loci in microsporidian phylogenetics is the absence of some orthologs in certain microsporidian lineages (e.g. loss of glycolytic genes in *Enterocytozoon bieneusi*) (Akiyoshi et al. 2009; Keeling et al. 2010). However, recent phylogenetic tools like RAXML (Stamatakis 2014) and MRBAYES are now able to take distinct gene evolution in a concatenated alignment into account. Studies have also shown that a multi-locus approach tolerates the absence of orthologs from individual lineages to a high degree (Delsuc et al. 2005) without affecting either the resolution or robustness of the final tree.

Presently, there is a compounding body of evidence supporting the idea that morphological and developmental features in the microsporidian phylum are indeed plastic between both closely and distantly related microsporidia (Vossbrinck & Debrunner-Vossbrinck 2005; Stentiford et al. 2013). Stentiford *et al.* (2013) observed that microsporidians isolated from marine decapod crustaceans that would have been classed as distantly related taxa (*Nadelspora* and *Ameson*) under a morphology-based classification system are close relatives on rDNA-based phylogenetic trees and are potential life cycle variants of the

same taxon (Stentiford et al. 2013). These authors together with Vossbrinck and Debrunner-Vossbrinck (2005) have also hinted to the widespread of this taxonomic misplacement within the entire microsporidian phylum.

If taxa have been truly misplaced across the microsporidian phylum, it could lead to serious consequences with respect to policy making for microsporidian infections in aquaculture. This is because taxonomic names of pathogenic species are fed into legislative frameworks that are used to inform policy making (Stentiford et al. 2014). For example, reporting the presence of a pathogenic species in the fisheries of a country could lead to an international embargo on the fish-host for fears of introduction of the pathogen into naïve waters. An example of this type of embargo is seen in the EC commission decision to safeguard UK waters from *Gyrodactylus salaris* [reviewed in (Stentiford et al. 2014)] by prohibiting the import of live salmonids into the UK (Fisheries Research Services). Since microsporidia are parasites of commercially important fisheries (Stentiford et al. 2013), the need to develop a more robust taxonomic framework to assist the accurate resolution of disease-causing taxa to the species level is urgent. Efforts are therefore currently being directed into complementing ultrastructural morphological data with rDNA-based phylogenetic evidence to reclassify previously erroneously placed taxa (Vossbrinck & Debrunner-Vossbrinck 2005; Brown & Adamson 2006; Stentiford et al. 2011).

In light of the advantages of multi-loci phylogenetics and the afore mentioned urgent need for a greater taxonomic resolution within the microsporidian phylum, this study presents a multi-gene phylogenetic analysis focusing on *Hepatospora eriocheir*, a parasite of the invasive Chinese mitten crab (Stentiford et al. 2011), a pea crab (*Pinnotheres pisum*) infecting microsporidium (Longshaw et al. 2012) and a novel microsporidium that infects commercially important edible crabs (*Cancer pagurus*). All three parasites infect the hepatopancreas of their respective hosts where heavy infection leads to the dissociation of epithelial cells from each other and subsequent sloughing of hepatopancreatic epithelium into the tubule lumen (Stentiford et al. 2011; Longshaw et al. 2012). These parasites also shared a similar life cycle that followed a merogony-plasmodium-sporogony-spore sequence, a characteristic feature of the Enterocytozoonidae family (Stentiford et al. 2011; Longshaw et al. 2012). These similarities transcended to the molecular level as sequencing and alignment of partial rDNA regions from the

three microsporidia performed by collaborators at CEFAS showed 99 % sequence identity.

Table 4.1: Morphological differences between *Hepatospora* and *Hepatospora*-like microsporidia.

	<i>H. eriocheir</i>	Pea crab parasite	Edible crab parasite
Karyotype	Unikaryotic	Dikaryotic	Dikaryotic
Polar tube arrangement	7-8 polar filament coils in a single rank	7-8 polar filament coils in a single rank	5-6 polar filament coils in a single rank

Despite these similarities, the three parasites differed in key morphological features presently used to classify microsporidia. For instance, *H. eriocheir* is unikaryotic whereas the pea crab and edible crab parasites are dikaryotic. There is also a noticeable difference in reported polar tube arrangement (see Table 4.1). Histopathology evidence also shows that these infections are detrimental to the animal's health as infected hosts are incapable of recovery (Longshaw et al. 2012). Since these infections occur in both invasive and commercially important crabs, this study aims to understand the phylogenetic position of the causative agents for the assignment of proper taxonomic names that could later be fed into legislative frameworks as mentioned earlier.

4.2 Main aims of study

To understand the phylogenetic relationship between *Hepatospora eriocheir*, a parasite of the invasive Chinese mitten crab (*Eriocheir sinensis*), a pea crab (*Pinnotheres pisum*) infecting microsporidium and a novel microsporidium that infects commercially important edible crabs (*Cancer pagurus*).

4.3 Methods

4.3.1 Sampling of edible crabs

See Section 2.3.1

4.3.2 Identification of infected tissues of the edible crab

See Section 2.3.2

4.3.3 Spore extraction from infected tissues isolated from the edible crab

See Section 2.3.3

4.3.4 Genomic DNA extraction

See Section 2.3.6 CEFAS collaborators used EZ1 tissue kits and BioRobots (Quiagen, UK) to extract total DNA from the pea crab parasite following manufacture's instructions. Dr. Bryony Williams of Exeter University, UK provided the assembled genome of *H. eriocheir*.

4.3.5 Illumina read assembly for the edible crab parasite

See Section 2.3.7

4.3.6 Marker genes used in this study

In this study, the following six marker genes were used to perform phylogenomic analyses as they had been successfully used in previous phylogenetic studies: amino acyl tRNA synthetases (Brown & Doolittle 1999): arginyl tRNA synthetase and prolyl tRNA synthetase, β -tubulin (Edlind et al. 1996), chitin synthase (Hinkle et al. 1997), heat shock protein 70, HSP70 (Hirt et al. 1997) and RNA polymerase II (Hirt et al. 1999).

4.3.7 Identification of six marker genes from assembled genomes of *Hepatospora eriocheir* and edible crab parasite

In order to identify the desired marker genes in the newly sequenced genome and that of *Hepatospora eriocheir*, their corresponding open reading frames (ORFs) identified with the GETORF bioinformatics tool (Rice et al. 2000) were initially converted into BLAST databases with the FORMATDB program (Mount 2007). *Vittaforma corneae* orthologs of the six marker genes were retrieved from the MICROSPORIDIADB online database (Aurrecoechea et al. 2011) and queried against the newly created ORF database with BLASTN (Mount 2007).

The top BLASTN hits were used to construct preliminary trees to assess their orthology.

4.3.8 Primer design, PCR and sequencing of six marker genes from the pea crab parasite

Due to the low amount of pea crab parasite genomic DNA collected by CEFAS collaborators, gene specific PCRs and subsequent sequencing was performed to retrieve the corresponding sequences for the six genes of the pea crab parasite rather than full genome sequencing. I designed the gene-specific primers by using the first and last 18 nucleotides of the selected orthologs from *H. eriocheir*. CEFAS collaborators performed subsequent PCR and sequencing.

4.3.8.1 Creation of a six-gene concatenated phylogenetic tree

4.3.8.1.1 Maximum likelihood analysis

The sequences of the six genes (amino acyl tRNA synthetases (Brown & Doolittle 1999): arginyl tRNA synthetase and prolyl tRNA synthetase, β -tubulin (Edlind et al. 1996), chitin synthase (Hinkle et al. 1997), heat shock protein 70, HSP70 (Hirt et al. 1997) and RNA polymerase II (Hirt et al. 1999)) amplified from the pea crab parasite were individually queried against the non-redundant MICROSPORIDIADB database (Aurrecochea et al. 2011) to identify orthologs for 20 other microsporidian species whose genomes are publicly available. The orthologs for each gene set were individually aligned with the MUSCLE program (v3.8.31) (Edgar 2004) using the default settings and then subsequently masked with the automatic command line tool TRIMAL (v1.2rev59) (Capella-Gutierrez et al. 2009). Homologs of *Saccharomyces cerevisiae* retrieved from SGD were used as an out-group in each of the microsporidian datasets.

A GTR substitution model with the GAMMA model of rate heterogeneity was used to construct maximum likelihood (ML) trees for the individual gene sets with the RAXML (v7.2.7) command line tool (Stamatakis 2014). These were pilot trees to check for unusually long-branch lengths indicative of unlikely orthologs. The masked genes from each microsporidian species were subsequently manually concatenated using the graphical interface of the SEAVIEW program (v4) (Gouy et al. 2010). For the final construction of the ML concatenated gene tree, a partition file that contained the positions of the individual genes within the alignment was manually created and passed to the RAXML program using the “-

q” option (Stamatakis 2014). This was to enable the program to treat each gene in the six-gene concatenated alignment separately and allow it to estimate individual nucleotide substitution rates for each of the six genes in the concatenated alignment. These estimations were also performed with the GTR+GAMMA nucleotide substitution model.

4.3.8.1.2 Bayesian inference analysis on six-gene concatenated alignment.

To check for reliability of the phylogenetic relationships estimated by maximum likelihood analysis, a Bayesian inference method was also used to reconstruct the six-gene concatenated phylogenetic tree using MRBAYES program (v3.2) (Ronquist et al. 2012). In order to take the different rates of nucleotide substitutions for individual genes into account, a partition file containing positions of the individual genes in the alignment was created according to the program manual. The program was run using a GTR+GAMMA model and probability distributions were generated using the Markov Chain Monte Carlo Methods. A total of 1,020,000 generations were run, the first 25 % of sampled trees were discarded as “burn-in” and a consensus tree was constructed.

4.4 Results

4.4.1 Creation of a six-gene concatenated phylogenetic tree

4.4.1.1 Individual gene trees

The selected microsporidian genes from the MICROSPORIDIADB database (Aurrecochea et al. 2011) did not show any unusual branch lengths on the individual phylogenetic trees thereby alluding to their identity as true orthologs (see Figure 4.1 A-F). Pea crab and Edible crab parasite orthologs always grouped together with those of *H. eriocheir*. Also, the *Hepatospora/Hepatospora*-like clade (see Figure 4.1 A-F shown in blue) always formed a sister group to *E. bieneusi*. Grouping was also observed for strains/subspecies of *Enc. cuniculi*, *Enc. hellem* and *Nematocida* spp. The *Nematocida* clade often formed the most basal group within the microsporidian phylum. *Nosema ceranae* and *Vittaforma corneae* consistently formed distinct separate groups and did not cluster with other microsporidian strains/ subspecies in the phylogenetic analyses of all 6 datasets.

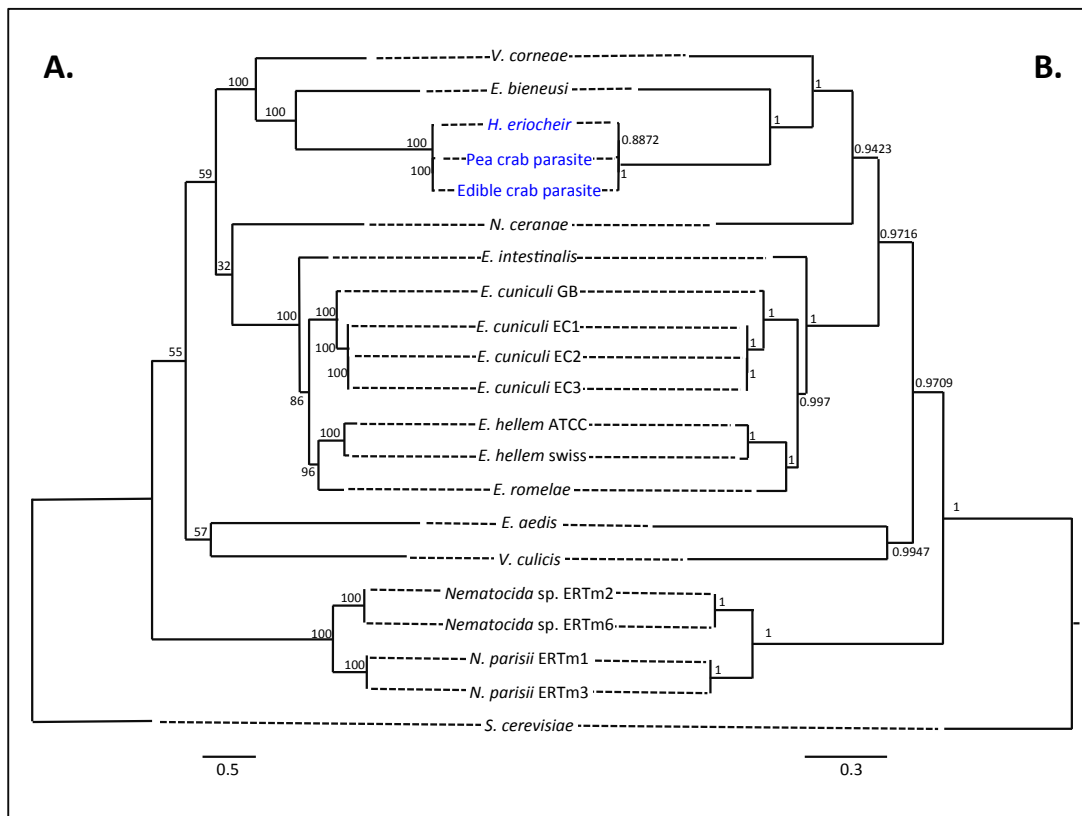


Figure 4.2: Phylogenetic trees based on A. maximum likelihood B. Bayesian inference of 20 microsporidians for six concatenated genes rooted with *Saccharomyces cerevisiae*. Numbers on nodes are A. Bootstrap confidence levels from 100 replicates B. Bayesian posterior probability values. Both trees displaying identical topologies and grouping of *Hepatospora/Hepatospora*-like clade shown in blue. The scale bars represent nucleotide substitutions per site.

Table 4.2: Higher sequence similarity between the three *Hepatospora/Hepatospora*-like species than between strains/subspecies of other microsporidia. The bottom left of matrix is percentage identity comparison of 18,323 sites (i.e. nucleotides+gaps) resulting from the alignment of six marker genes of three *Hepatospora/Hepatospora*-like microsporidia and 18 other microsporidian species. Top right of matrix represent number of variable nucleotides in the pairwise alignment of two species without taking gaps into account (number given of variable nucleotides/total number of aligned nucleotides).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. <i>N. parisii</i> ERTm1	-	14/ 12657																			
2. <i>N. parisii</i> ERTm3	99.77	-																			
3. <i>Nematocida</i> sp. ERTm2	78.79	78.79	-	148/ 12543																	
4. <i>Nematocida</i> sp. ERTm6	78.86	78.84	98.65	-																	
5. <i>Vav. culicis</i>	51.75	51.78	51.65	51.95	-																
6. <i>S. cerevisiae</i>	50.23	50.26	49.65	49.56	49.91	-															
7. <i>Enc. cuniculi</i> EC2	53.72	53.75	53.89	54.14	55.44	52.97	-	6/ 12447	854/ 12228	30/ 12447											
8. <i>Enc. cuniculi</i> EC3	53.72	53.75	53.90	54.14	55.45	52.94	99.95	-	851/ 12228	28/ 12447											
9. <i>Enc. cuniculi</i> GB	54.49	54.51	54.57	54.84	54.86	54.21	92.10	92.13	-	825/ 12228											
10. <i>Enc. cuniculi</i> EC1	53.70	53.74	53.87	54.12	55.51	52.94	99.75	99.77	92.34	-											
11. <i>Enc. intestinalis</i>	56.57	56.59	56.20	56.64	56.61	55.64	79.59	79.59	81.06	79.56	-										
12. <i>Enc. romaleae</i>	53.63	53.62	52.72	53.07	51.74	54.38	74.76	74.78	79.69	74.78	71.10	-									
13. <i>Enc. hellem</i> ATCC	55.07	55.11	55.08	55.55	55.91	53.89	78.04	78.05	77.86	77.99	78.85	73.75	-	715/ 11472							
14. <i>Enc. hellem</i> Swiss	53.84	53.84	53.93	54.27	55.45	53.35	82.02	82.04	76.52	82.02	79.54	78.14	93.22	-							
15. <i>N. ceranae</i>	54.08	54.10	52.52	53.02	52.59	56.21	60.57	60.57	62.05	60.56	57.33	51.57	57.73	61.57	-						
16. <i>V. corneae</i>	53.87	53.86	54.18	54.46	53.50	53.70	60.74	60.76	61.42	60.71	60	55.81	58.53	60.38	49.68	-					
17. <i>E. bieneusi</i>	49.23	49.26	48.14	48.52	48.16	51.87	53.54	53.55	54.67	53.54	53.62	50.93	51.53	54.29	56.56	53.31	-				
18. <i>Edh. aedis</i>	44.59	44.64	44.21	44.48	42.74	43.12	46.20	46.19	45.53	46.26	45.25	46.12	43.95	46.55	42.65	43.16	39.84	-			
19. <i>H. eriocheir</i>	53.53	53.52	51.44	51.74	53.09	56.49	57.16	57.19	57.78	57.19	58.72	46.01	57.13	58.08	53.02	52.46	57.59	43.43	-	134/ 9927	44/ 3668
20. Edible crab parasite	51.06	51.06	49.41	49.67	50.91	54.30	54.85	54.88	56.10	54.87	58.96	57.40	56.55	55.70	61.60	60.20	59.49	45.13	98.52	-	41/ 3668
21. Pea crab parasite	55.58	55.61	54.74	54.55	55.07	58.47	58.58	58.58	59.40	58.61	60.52	55.23	58.04	58.94	60.03	56.92	65.75	36.96	98.80	98.88	-

4.4.1.2 Six-gene concatenated tree

The resulting alignment of the concatenated 6 genes after trimming uninformative sites consisted of 18,232 sites. Phylogenetic trees based on ML and BI methods displayed identical topologies and strongly supported the grouping of *Hepatospora/Hepatospora*-like parasites, *Enc. cuniculi* strains and *Nematocida* spp. as distinct clades with high confidence values (see Figure 4.2). The *Hepatospora/Hepatospora*-like parasite clade branched as a sister group to *E. bieneusi*. As characterised by the topologies of the individual gene trees, the *Nematocida* clade formed the most basal group within the microsporidian phylum.

4.5 Discussion

4.5.1 Taxonomic names for the edible and pea crab parasites

The rationale behind the multi-gene phylogeny approach was to take advantage of the increased number of characters available from using genome projects to resolve the branching relationship between the three *Hepatospora/Hepatospora*-like microsporidia. This study resulted in identical tree topologies in both the ML and BI probabilistic approaches used with nodes supported by high bootstrap and posterior probability values respectively. This, in addition to the retrieval of well known relationships like the grouping of *Encephalitozoon* and *Nematocida* subspecies (see Figure 4.2) and overall tree topology similarity to previously published work (Troemel et al. 2008; Stentiford et al. 2011; Cuomo et al. 2012) increases confidence in the phylogenetic relationships inferred by this study.

The placement of *Hepatospora/Hepatospora*-like parasites as a sister group to *E. bieneusi* is not surprising as microscopy analysis shows that these parasites undergo pre-sporogonial plasmodia formation, which is characteristic of members of the Enterocytozoonidae family of which *E. bieneusi* forms part (Stentiford et al. 2011; Longshaw et al. 2012)[Bateman Pers. Comm.]. Stentiford et al. (2011) erected the *Hepatospora* genus in 2011 to encompass hepatopancreas-infecting microsporidia (Stentiford et al. 2011). Since microscopy evidence suggests that the pea crab and edible crab parasites undergo their entire life cycle within the hepatopancreas of their respective decapod hosts (Longshaw et al. 2012) [Bateman, pers. comm.], the addition of these parasites to the *Hepatospora* genus is justified. Considering the high nucleotide sequence similarity between the three *Hepatospora* parasites as compared to other closely related microsporidia (see highlighted boxes in Table 4.2) and their consequent grouping with minimal branching distances (shown in blue in Figure 4.2), it is likely that these parasites are in fact subspecies of *H. eriocheir*. I therefore propose the assignment of *Hepatospora eriocheir pinnotheres* and *Hepatospora eriocheir canceri* subspecies names to the pea crab and edible crab parasites respectively.

At least 100 crabs were sampled in each of the 3 surveillance studies in which these *Hepatospora* subspecies were discovered (Stentiford et al. 2011; Longshaw et al. 2012)[Bateman, pers. comm.]. There was approximately 70 %, 5 % and 1 % *Hepatospora* subspecies prevalence in the sampled Chinese mitten

crab, pea crab and edible crab respectively. Microscopy evidence in all three cases showed no instance of co-infection of these *Hepatospora* subspecies within the same host (Stentiford et al. 2011; Longshaw et al. 2012), [Bateman, pers. comm.] ruling out the possibility for these three parasites to be different morphological life stages of each other. This absence of co-infections is noteworthy as the hosts in all three cases are decapods that share ecological marine water niches at least at some point in their life cycle (Becker & Türkay 2010; Clark et al. 1998; Ingle 1980). The fact that the edible crab could be a potential predator of the other two decapods hosts (Lawton 1989) and thereby increase the chance of inter-host transmission makes the absence of co-infections even more fascinating. This study therefore presents an example of strong host specificity of decapod-infecting microsporidia towards their respective hosts despite 99 % DNA similarity between them (see blue highlighted boxes in table 1.2).

4.5.2 Importance of continuous disease profile surveillance for UK fisheries

Despite the strong host specificity exhibited by these *Hepatospora* subspecies, microsporidians are well known to have the ability to adapt to different host environments. For example *Enc. cuniculi* strain III causes asymptomatic infections in dogs but has been continuously reported and sometimes linked to fatalities in humans (Orenstein et al. 2005; Snowden et al. 1999; Didier et al. 1996; Reetz et al. 2009). An extreme example where infections may have transferred across host phyla is seen in Slodkowicz *et al.*'s study where they identified a high occurrence of human infecting microsporidians in avian species (Slodkowicz-Kowalska et al. 2006). Considering these cases of microsporidian transmission across both host genus and phyla, it is important that disease profiles of these UK decapods and their economically important ecological neighbours like velvet swimming crabs, lobsters and *Nephrops* are continuously monitored as a cross over to immunologically naïve hosts could have devastating effects.

4.5.3 Origin of *Hepatospora eriocheir*

As their name implies, the Chinese mitten crab is an invasive UK species originating from Eastern China (Clark et al. 1998). This crab has been reported to have a long-range aquatic and terrestrial migration ability and a possibility to

serve as a host for other parasites in its native land (Clark et al. 1998). For this reason, there is the tendency for the mitten crab to serve as a carrier of foreign pathogens that could be harmful to local species in the UK.

The lower cases of infections in UK crabs could also arguably be due to difficulty of *Hepatospora* parasites to adapt to a new host environment. However, this would have probably resulted in the occurrence of infections predominantly in stressed UK crabs. Interestingly, there was no observable difference in the morphological appearance or physiology between infected and non-infected crabs used in these studies (Stentiford et al. 2011; Longshaw et al. 2012)[Bateman, pers. comm.] suggesting that this may not be the case. Another reason why it is currently difficult to link *Hepatospora* infections in UK species to a Chinese origin is because phylogenetic studies to conclusively link *Hepatospora eriocheir* found in native Chinese mitten crabs to invasive Chinese mitten crabs found in the UK still remains to be undertaken.

All three crab hosts species infected with *Hepatospora eriocheir* subspecies belong to the Eubrachyuran family (Tsang et al. 2014). This could imply a single *Hepatospora eriocheir* introduction in the primordial Eubrachyuran crab prior to the divergence of fresh water crabs which dates back to about 150 million years ago (Tsang et al. 2014) and the evolution of these *Hepatospora eriocheir* subspecies in parallel with their respective crab hosts explaining their host specificity. This could also suggest a ubiquitous occurrence of *Hepatospora eriocheir* infections in Eubrachyuran crab species. Considering that the microsporidian phylum may have diverged more than 600 million years ago (Lücking et al. 2009), the introduction of *Hepatospora eriocheir* into crab hosts occurred comparatively recently in the history of microsporidian evolution explaining the observable genomic DNA similarity despite morphological divergence.

4.5.4 Limitations of study and future outlook

A major drawback of this study is the use of highly conserved genes in this analysis. Even though these six marker genes and other protein coding genes have been successfully used in previous studies to infer deep phylogenetic relationships within the microsporidian phylum and placing microsporidians within the tree of life (Hirt et al. 1999; Keeling 2003; Fast et al. 1999; Vivarès et al. 2002), these genes were not able to properly resolve branching relationships between

the three *Hepatospora* subspecies in this analysis (see Figure 4.2). This was due to their high level of nucleotide sequence similarity between the *Hepatospora* subspecies.

Due to the small sizes of pea crab hepatopancreatic material and minimal *Hepatospora eriocheir pinnotheres* infection levels, colleagues at CEFAS were only able to extract a very small quantity of parasite genomic DNA. This in turn limited the number of marker genes that could be amplified from the pea crab parasite for this study. Since polar tube and spore wall protein coding genes have been used in previous studies to distinguish between *Encephalitozoon* strains (Peuvel et al. 2000; Polonais et al. 2010), these proteins could have been used in these analyses to improve the resolution of branching relationship between the *Hepatospora* subspecies if not for the shortage of genomic DNA material. It must however be noted that even though the above mentioned studies were successful in differentiating between *Encephalitozoon* strains by looking at nucleotide polymorphisms of polar tube protein coding genes, this gene was unsuccessful in differentiating between *Nosema* species in similar recent studies (Roudel et al. 2013; Van der Zee et al. 2014).

This probably highlights the varying evolutionary pressures that the same genes in different microsporidian lineages are subject to and that an orthologous gene that resolves close phylogenetic relationships in a specific microsporidia lineage may not necessarily do so in another lineage. Future studies of this kind should probably focus on the most divergent single copy orthologs between strains/subspecies of interest. However, identifying orthologs across the entire microsporidian phylum for genes that are divergent even between closely related species would prove problematic due to general low level of similarity (~31 %) between core genes of members of this phylum (Cuomo et al. 2012).

A two-step supertree construction could perhaps be the way forward to tackle this problem in future studies. Here a general microsporidian phylogenetic tree is first constructed with concatenated genes that are conserved across the phylum to correctly place the individual internal clades. Individual trees are subsequently constructed for clades of closely related species with a concatenated set of clade specific single copy orthologous genes that are highly divergent. The first tree is finally merged with the individual trees of closely related species to create a supertree. This approach will therefore correctly place clades of microsporidian species within the phylum but also resolve branching relationships between

closely related species. This concept of supertrees has been around since 1986 [reviewed in (Bininda-Emonds 2004)]. However this approach have been used to focus on higher taxonomic levels (Bininda-Emonds et al. 2002) rather than lower taxonomic levels as I propose.

In summary, there is a no visible difference in branching relationships between closely related species and overall tree topology when results from this study are compared to previous rDNA based phylogenies (Freeman et al. 2013). These results also support the clustering of all three crab-infecting *Hepatospora*-like organisms and their positioning as a sister group to human-infecting *E. bieneusi*. As microsporidian genomes become increasingly available, a switch from phylogenetics to phylogenomics is likely as the latter presents a more holistic approach to understanding close phylogenetic relationships and providing information for more robust taxonomic assignments.

4.6 Conclusion

These results are in concordance with earlier results from the rDNA data and therefore reinforce findings of previous authors (Silveira & Canning; Sokolova et al. 2009) that microsporidian phylogeny is poorly resolved by morphology, even at species level. In light of morphological studies carried out by colleagues at CEFAS (Stentiford et al. 2011; Longshaw et al. 2012) [Bateman Pers. Comm.] and results from this study suggesting a close phylogenetic relationship between the three *Hepatospora/Hepatospora*-like parasites, I propose these parasites be considered as the same species despite reported difference in karyotype and spore morphology. I suggest the assignment of subspecies names: *Hepatospora eriocheir pinnotheres* and *Hepatospora eriocheir canceri* to the pea crab and edible crab parasite respectively. Analyses performed in this chapter have been published in a co-authored paper, Bateman *et al.* 2016 (See Appendix 12).

Chapter 5 Summary and future perspectives

5.1 New genomic data for four Enterocytozoonidae species: Impact on aquaculture and global food security

One of the reasons for studying the Enterocytozoonidae is that they are a major threat to the global aquaculture industry (Stentiford, Feist, et al. 2013; Stentiford et al. 2016). Member species of this microsporidian family cause serious annual yield loss in farmed fisheries that ultimately result in the loss of millions of pounds in revenue (Desportes et al. 1985; Hedrick et al. 1991; Chilmonczyk et al. 1991; Lom & Dykoá 2002; Stentiford & Bateman 2007; Stentiford et al. 2007; Tourtip et al. 2009; Freeman & Sommerville 2009; Nylund et al. 2010; Stentiford et al. 2011; Freeman et al. 2013) (Table 1.1). Aquaculture has been flagged as a potential avenue to be exploited in order to meet the extra food demand associated with the ever-increasing global population (Duarte et al. 2009). However, aquaculture production will need to increase significantly within the next 30 years in order to meet forecasted global demand (Kearney 2010).

A past lesson learnt with regards to Enterocytozoonidae infections is that intense farming of aquatic animals is often accompanied by increased prevalence to diseases associated with microsporidian infections. An example is the recent increased occurrence of *Enterocytozoon hepatopenaei* infections in Asian farmed shrimp species. This has been associated with Early Mortality Syndrome, a disease currently plaguing the Asian shrimp Industry (Stentiford et al. 2016). Another example is the increased prevalence of *Enterospora nucleophila* infections in farmed juvenile gilthead sea bream. These infections have been linked to the Winter Disease Syndrome (Doimi 1996; Tort, Rotllant, et al. 1998; Domenech et al. 1999). Finally, *Hepatospora eriocheir* infections have been reported in invasive Chinese mitten crabs in the UK. Sampled crabs were living in the wild and although histopathology assessment showed heavy necrosis of hepatopancreatic tissues, the infected animal did not display any external symptoms of the disease (Stentiford et al. 2011). Contrastingly, farmed native Chinese Mitten crabs infected with the same parasite displayed severe impairment to their locomotive abilities and 40 % mortality rate (Ding et al. 2015). These examples indicate that current intensive aquaculture settings facilitate microsporidiosis in farmed aquatic animals. Some authors have suggested host proximity to be a reason for this (Stentiford et al. 2016). Regardless, it evident

that microsporidian-related disease in aquatic animals are a bottleneck that limit the realisation of aquaculture's full potential as a food source. As such, it is imperative that therapeutic or prophylactic measures against etiological agents such as the Enterocytozoonidae species are developed if aquaculture is truly going to be a candidate solution for future food security.

Many described microsporidian species are refractive to *in vitro* culturing and no technique has yet been developed for the molecular manipulation of these parasites. This has limited research in the development of novel therapeutic and prophylactic strategies to combat disease caused by microsporidian infections. Recent cheap Next Generation Sequencing and increased accessibility to high performance computing has however offered an alternative avenue for research into measures to combat disease caused by microsporidian species that are refractive to *in vitro* culturing such as members of the Enterocytozoonidae.

To this end this study has provided genomic data for four members of the Enterocytozoonidae family, namely *Enterocytozoon hepatopenaei*, *Enterospora canceri*, *Hepatospora eriocheir* and *Hepatospora eriocheir canceri*. By performing comparative analysis between these genomes and that of publicly available species [includes *Enterocytozoon bieneusi*, 18 other non-Enterocytozoonidae species and the genome of *Thelohania* sp. that was sequenced in this study], this study has identified a set of putative plasma membrane transporter proteins unique to members of the Enterocytozoonidae. Furthermore the genomic data produced by this study has provided a list of candidate effector proteins that may be secreted by these parasites to manipulate the host cell. Since plasma membrane transporters are exposed to the outside of the cell, epitopes of these proteins could be exploited to develop vaccines to offer immunity to the farmed fisheries affected by diseases caused by member species of the Enterocytozoonidae. Although, crustaceans lack the classic adaptive immunity that is present in jawed vertebrates (Söderhäll & Thörnqvist 1997), there is a growing body of evidence hinting to the presence of an alternate immune mechanism that can be primed against pathogenic infections with protein particles from pathogens (Yang et al. 2012; Valdez et al. 2014; Jia et al. 2016; Solís-Lucero et al. 2016). Sequences of putative secreted proteins identified in this study could be used in Yeast two-hybrid assays to identify interactions between parasite and host proteins. Results from such experiments could help in elucidating how these microsporidian parasites manipulate the host environment

and perhaps aid in engineering host strains less susceptible to diseases caused by member species of the Enterocytozoonidae.

5.2 Systematics of the Enterocytozoonidae

Microsporidian systematics has long been based on ultrastructural morphology and karyotypic evidence (Canning 1953; Vávra & Undeen 1970; Shadduck et al. 1990; Cali et al. 1993). More recent studies have harnessed rDNA sequences (Vossbrinck et al. 1998; Vossbrinck & Debrunner-Vossbrinck 2005). Although both strategies have significantly contributed to microsporidian systematics, there are problems inherent to these techniques. For instance emerging data suggests that the morphology of microsporidian spores is plastic and hence an inadequate defining factor for taxonomic ranking (Stentiford, Bateman, et al. 2013). Secondly, microsporidian rDNA sequences are not homogenous in some species and this has led authors to suggest the use of protein coding sequences (ideally, multiple protein coding sequences) for the purposes of inferring phylogenetic relationships between closely related species. Until this study, inference of evolutionary relationship between members of the Enterocytozoonidae via a multi-protein approach was impossible as *E. bienewisi*'s genome was the only one publicly available.

In this study, the evolutionary relationship between *E. bienewisi* and four members of this family (named above) was inferred via a phylogenomic approach. Phylogenomic results based on 21 protein coding sequences demonstrates that *Ent. canceri* and *E. hepatopenaei* have a closer evolutionary relationship than *E. bienewisi* and *E. hepatopenaei*. These results imply that the shrimp parasite, *E. hepatopenaei* should perhaps be removed from the *Enterocytozoon* genus and be assigned to the *Enterospora*. The separate phylogenomic analysis performed with six protein coding genes revealed that the *Hepatospora*-like organisms, isolated from different crab hosts are in fact the same species despite their varying morphological traits. Systematics within the Enterocytozoonidae family is a non-trivial subject as taxonomic names arising from phylogenetic studies such as this could be fed into legislative frameworks used to inform policies to mitigate the spread of diseases caused by these parasites (Stentiford et al. 2014).

Identifying proteins that are conserved across the entire microsporidian phylum to design primers and yet possess enough variation within their amino acid/nucleotide sequence is a difficult if not impossible task due to the

accelerated rate of evolution characterised by microsporidian genomes (Nakjang et al. 2013). Future studies aimed at distinguishing between closely related species should employ supertrees: Here, protein-coding sequences that are broadly conserved across genomes are used to infer distant evolutionary relationships. Nested trees are constructed with proteins that are perhaps unique to the closely related species of interest but contain a certain level of sequence variation. Although hypothetical, this idea should not be too difficult to implement considering computational advances of recent years.

5.3 Loss of glycolysis: a common trait within the Enterocytozoonidae

Data presented in this thesis demonstrate that absence of glycolysis is not a feature unique to *E. bienersi* but a common feature within the Enterocytozoonidae. This loss was not as a result of a single event in the common ancestor of the Enterocytozoonidae but instead due to relaxation of selective pressures to keep this pathway at the base in the common ancestor of the group. Despite the differential loss of glycolytic genes within this group, hexokinase was retained in each of the surveyed genomes. Results presented in this study hints to possible sensitivity of *Ent. canceri*'s chimeric hexokinase to glucose although this finding needs to be corroborated by further research. Regardless of the substrate sensitivity of the hexokinases coded by these aglycolytic genomes, the absence of ATP-releasing glycolytic enzymes such as pyruvate kinase in all these genomes leaves the enigma of how they activate their spores to invade host cells in the absence of their own ATP generation system. It is possible that these aglycolytic parasites are taken up by their host cells via phagocytosis and germinate their spores within the host cells by using host ATP via ATP/ADP translocases situated on spore surfaces. For the intranuclear parasite, *Ent. canceri* the polar filament may be specifically targeted to the host nucleus. This hypotheses need to be further investigated, perhaps starting with transcriptomic analysis of spores ingested by the crab host to see if ATP/ADP translocases are transcribed during this stage.

5.4 Intranuclear living may have led to gene dosage increase of plasma membrane transporters

One of the aims of this study was to understand how the intranuclear parasite, *Ent. canceri* obtained energy within its unusual environment and without access to the host mitochondria. There is evidence to suggest that there is no difference

in cytosolic and nuclear ATP levels implying that adaptation to intranuclear living does not put the parasite at a disadvantage with regards to access to host ATP (Imamura et al. 2009). Comparative genomic analysis performed here agrees with previous findings showing that microsporidian genomes do not possess a full gene set for *de novo* deoxyribonucleotide synthesis (Katinka et al. 2001; Heinz et al. 2012; Cuomo et al. 2012). In light of this, the most parsimonious explanation for the evolution of intranuclear survival is for increased accessibility to these DNA building blocks for rapid proliferation during merogony. Although results presented here suggest that intranuclear living has led to the increase of transporter coding genes, none of the transporters identified were unique to the intranuclear parasite. That is all transporter-coding genes identified in the genome *Ent. canceri* possessed an ortholog in at least one of the other genomes of non-intranuclear microsporidians analysed. This implies that intranuclear living did not require the evolution of novel plasma membrane transporters.

Genomes presented in this study are by no means the smallest microsporidian genomes described to date but are clearly the most metabolically reduced described yet. This work is a foundation for possible future work in understanding minimal eukaryotic evolution, developing therapeutic strategies to combat early mortality syndrome and understanding strategies employed by parasites to survive in unusual intracellular niches such as the nucleus.

Appendix 1: Assessing the taxonomic profile of 381-core-eukaryotic proteins present in the assembled genome of *Enterocytozoon hepatopenaei*

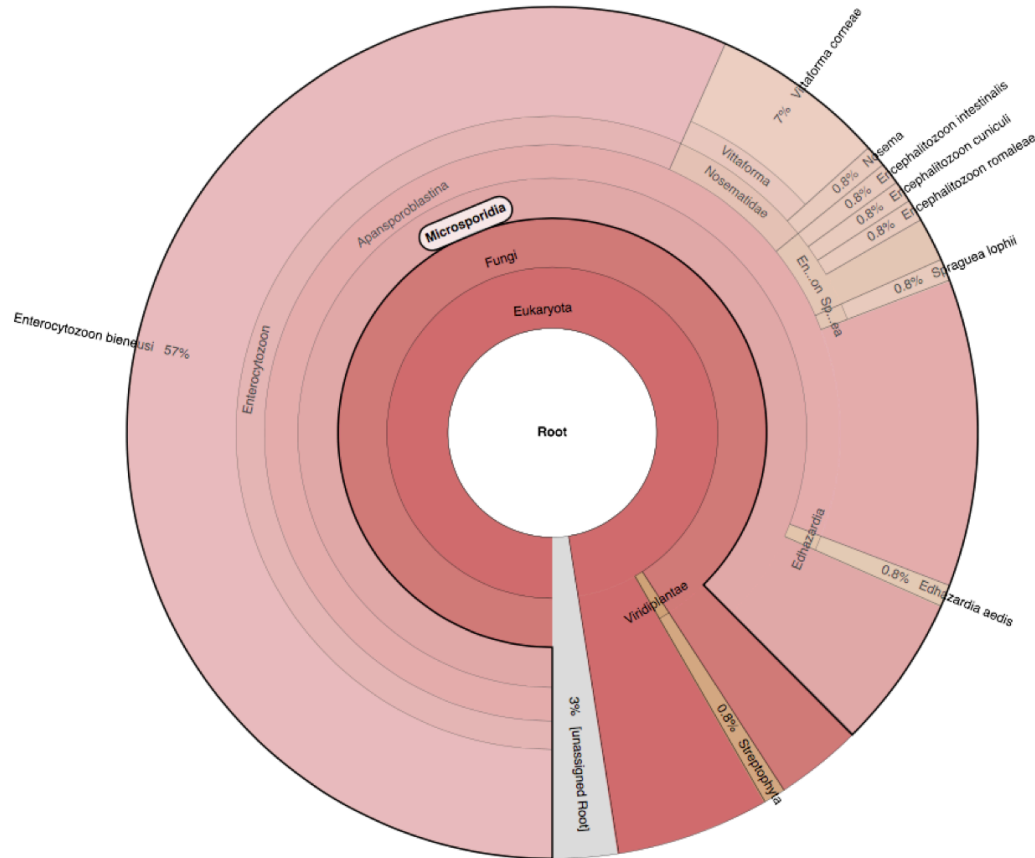


Figure 1. Taxonomic profile of the 120/381 proteins present in the *Enterocytozoon hepatopenaei* genome assembled with Illumina reads filtered at GC 24 %. Percentages refer to the proportion of the 120 proteins predicted by KRONA to belong to a specific taxonomic group.

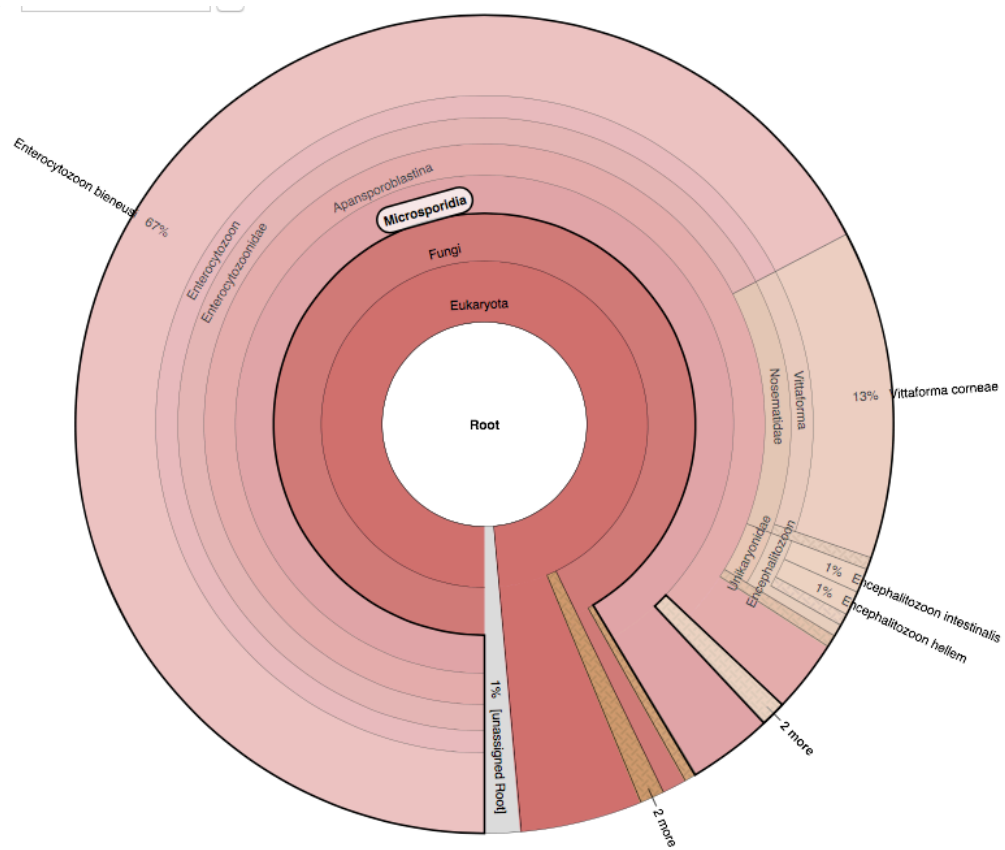


Figure 2. Taxonomic profile of the 208/381 proteins present in the *Enterocytozoon hepatopenaei* genome assembled with Illumina reads filtered at GC 27 %. Percentages refer to the proportion of the 208 proteins predicted by KRONA to belong to a specific taxonomic group.

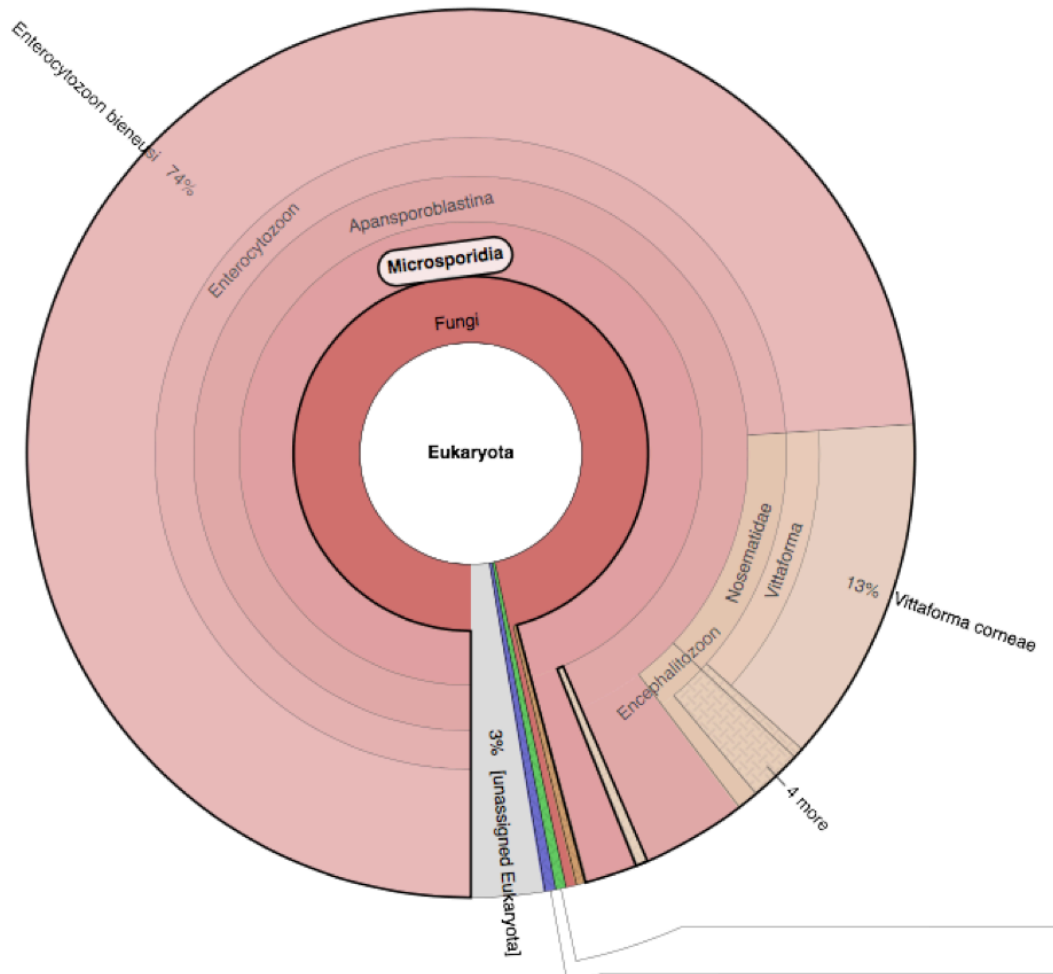


Figure 2. Taxonomic profile of the 261/381 proteins present in the *Enterocytozoon hepatopenaei* genome assembled with Illumina reads filtered at GC 30 %. Percentages refer to the proportion of the 261 proteins predicted by KRONA to belong to a specific taxonomic group. Green and blue belong to Viridiplantae and Neoptera respectively.

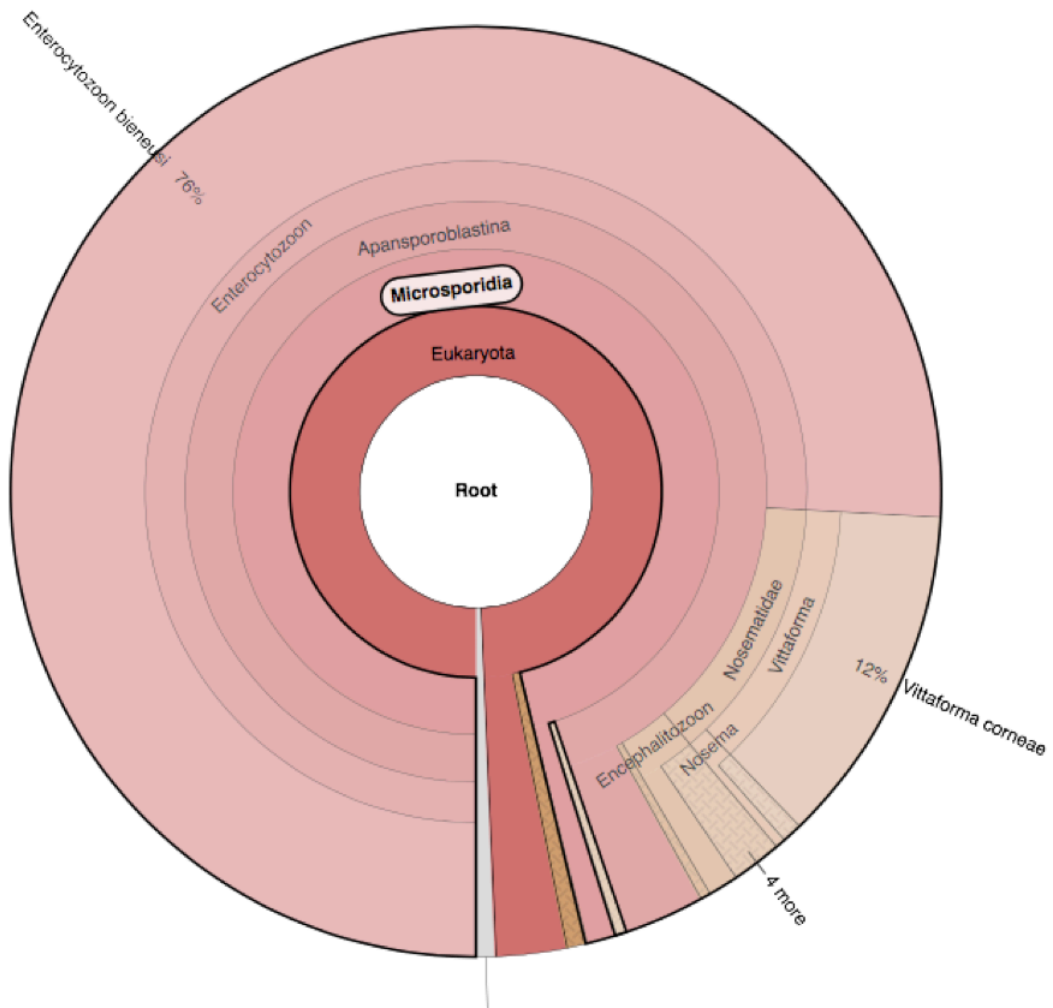


Figure 3. Taxonomic profile of the 286/381 proteins present in the *Enterocytozoon hepatopenaei* genome assembled with Illumina reads filtered at GC 33 %. Percentages refer to the proportion of the 286 proteins predicted by KRONA to belong to a specific taxonomic group.

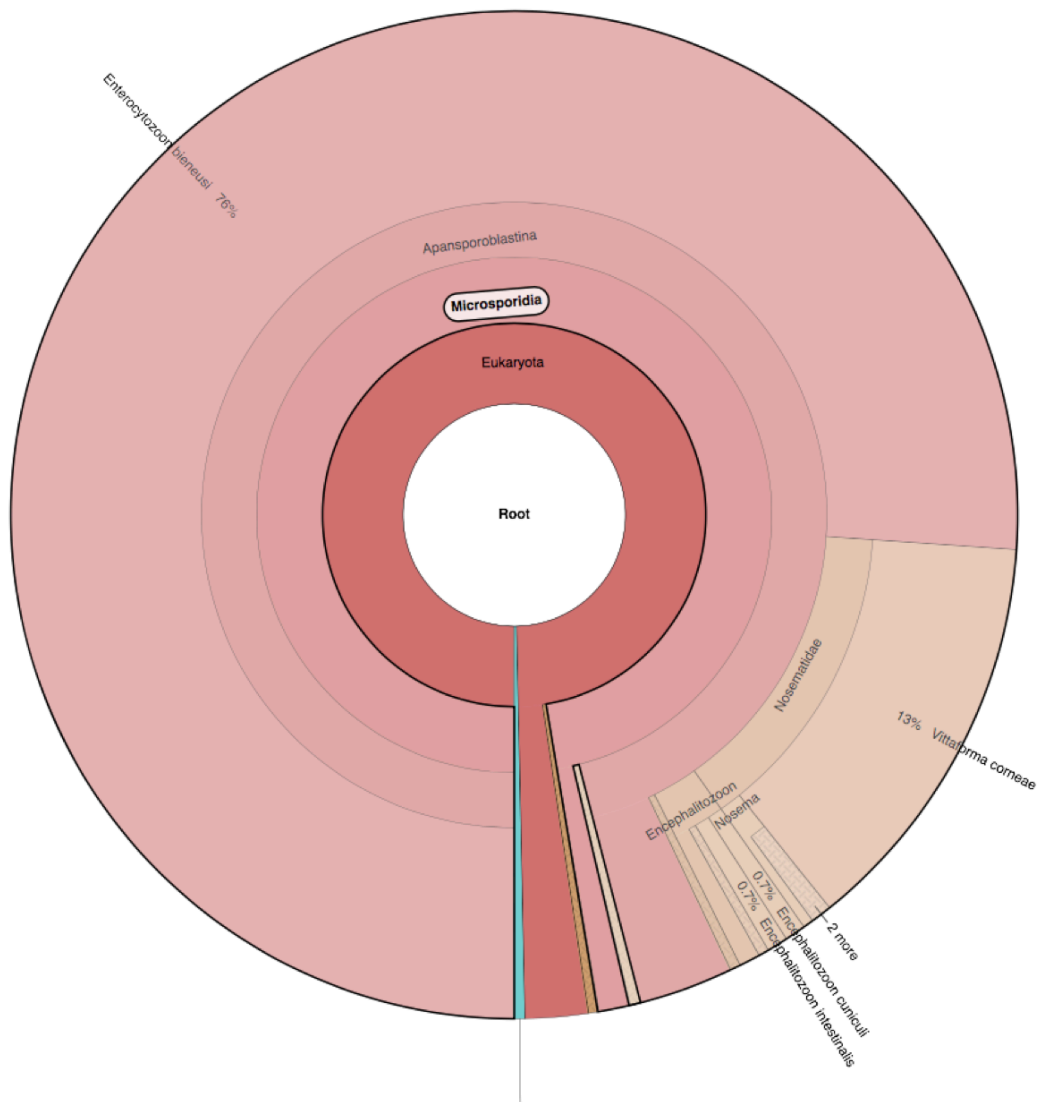


Figure 5: Taxonomic profile of the 295/381 proteins present in the *Enterocytozoon hepatopenaei* genome assembled with Illumina reads filtered at GC 36 %. Percentages refer to the proportion of the 295 proteins predicted by KRONA to belong to a specific taxonomic group. Blue section belongs to Gammaproteobacteria.

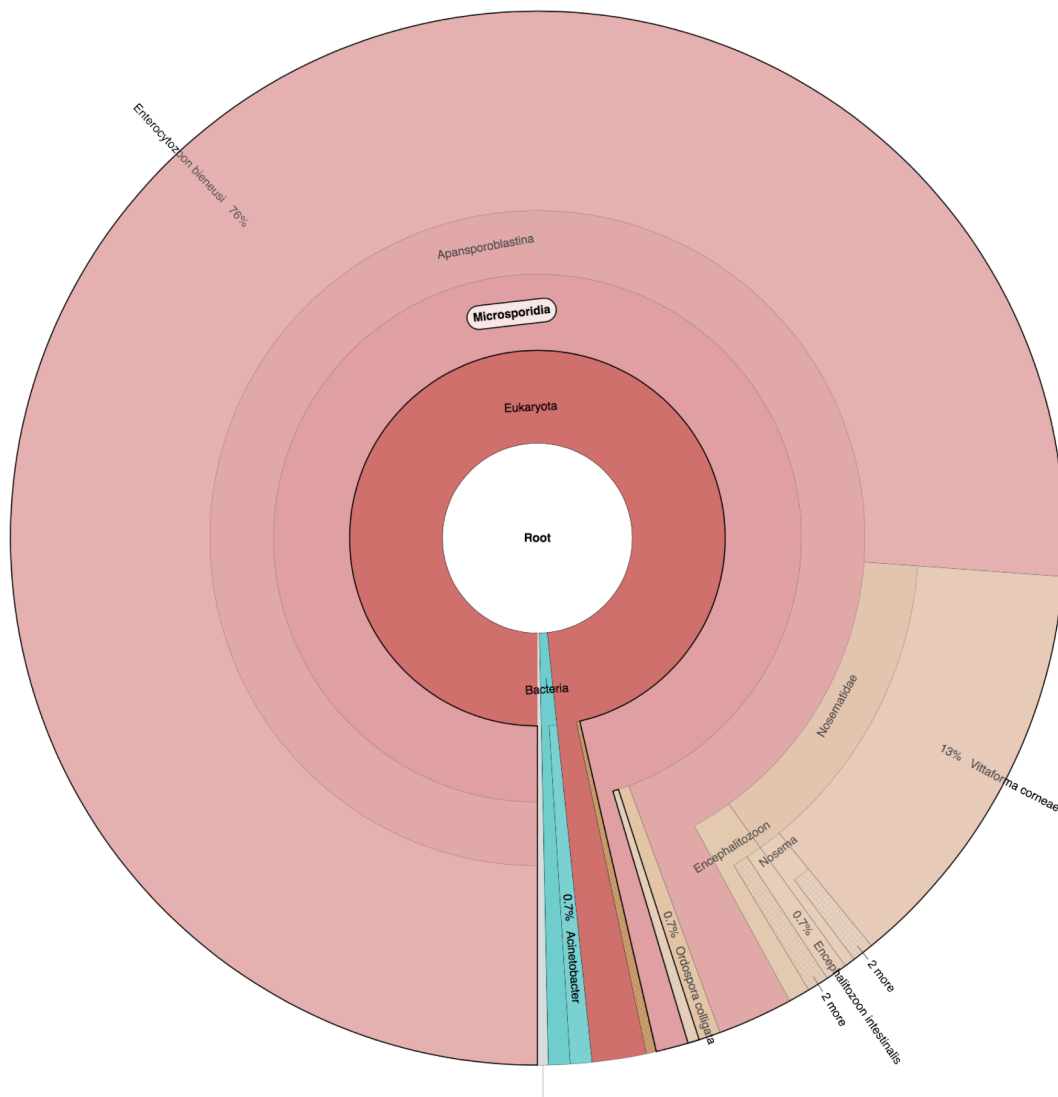


Figure 6. Taxonomic profile of the 302/381 proteins present in the *Enterocytozoon hepatopenaei* genome assembled with Illumina reads filtered at GC 39 %. Percentages refer to the proportion of the 302 proteins predicted by KRONA to belong to a specific taxonomic group.

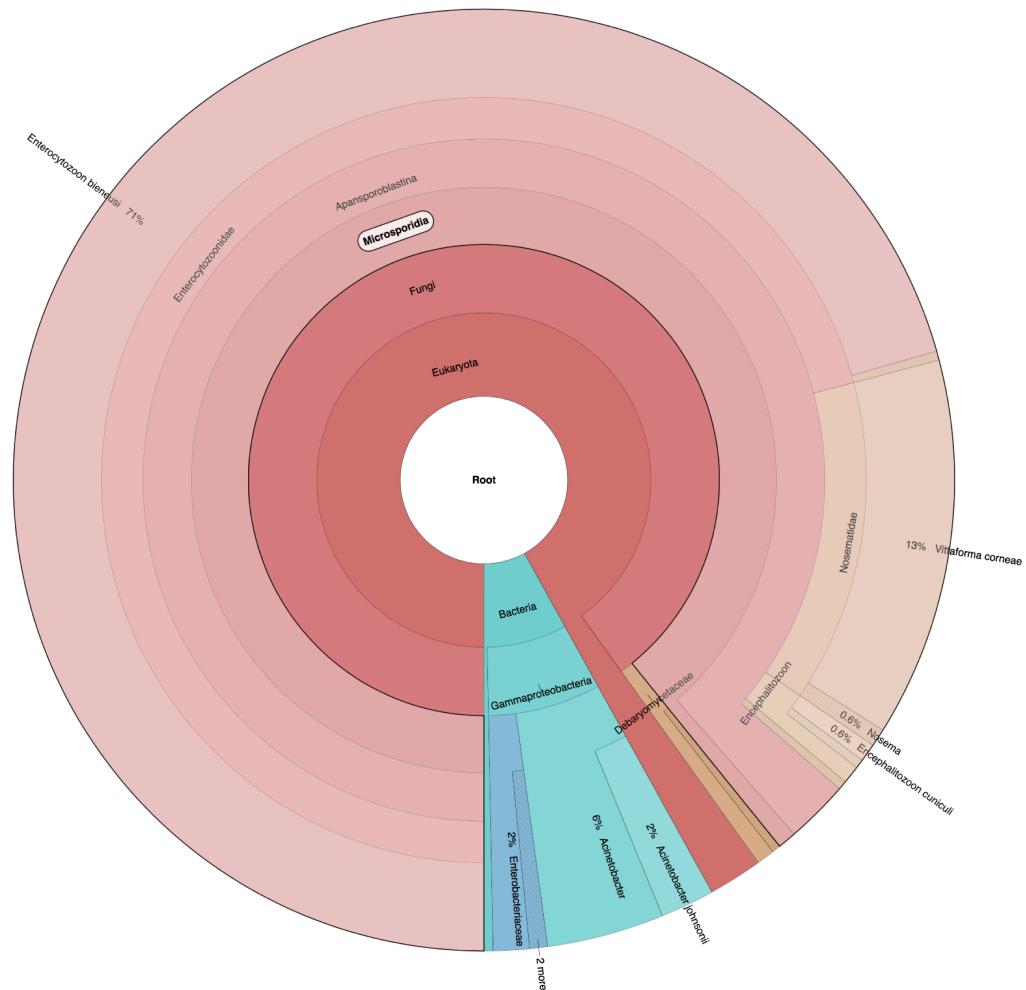


Figure 7. Taxonomic profile of the 323/381 proteins present in the *Enterocytozoon hepatopenaei* genome assembled with unfiltered Illumina reads. Percentages refer to the proportion of the 302 proteins predicted by KRONA to belong to a specific taxonomic group.

Appendix 2: Plasmids used in this study

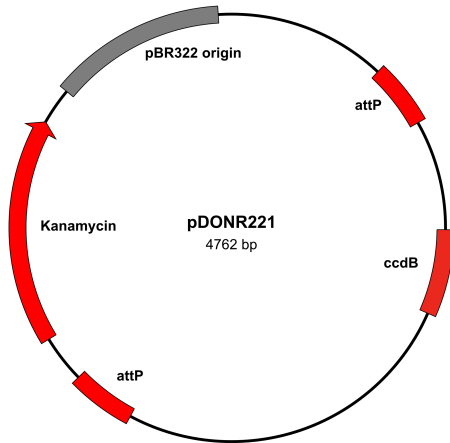


Figure 1: Schematic representation of the pDONR221 Gateway plasmid kindly donated by Graham Thomas, University of Exeter.

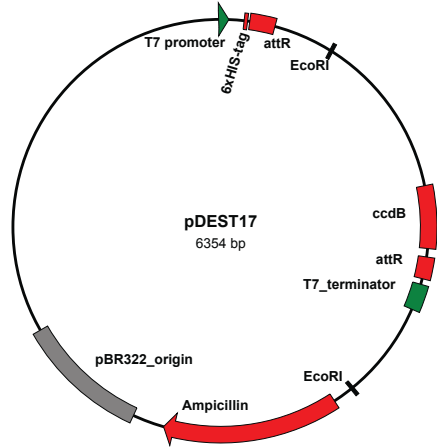


Figure 2: Schematic representation of the pDEST17 Gateway plasmid kindly donated by Graham Thomas, University of Exeter.

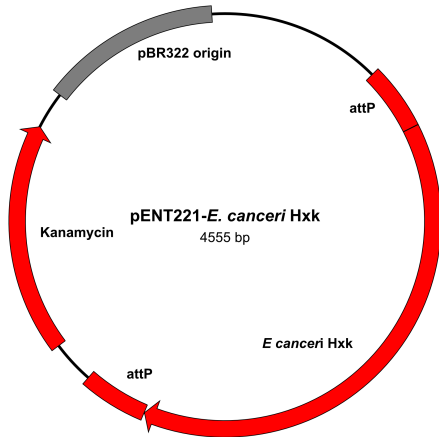


Figure 3: Schematic representation of the pENT221-*Ent. canceri* Hxk clone resulting from BP cloning between pDONR221 and *Ent. canceri* Hxk second PCR product.

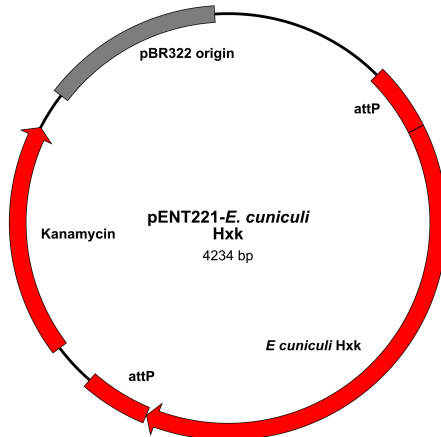


Figure 4: Schematic representation of the pENT221-*Enc. cuniculi* Hxk clone resulting from BP cloning between pDONR221 and *Enc. cuniculi* Hxk second PCR product.

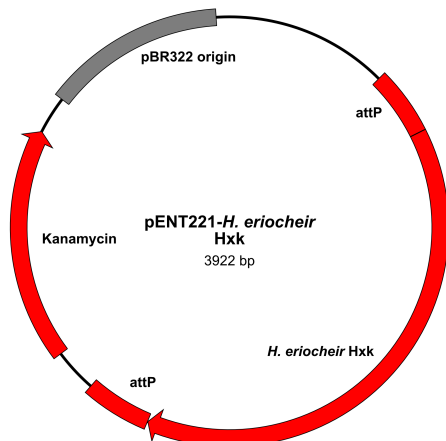


Figure 5: Schematic representation of the pENT221-*H. eriocheir* Hxk clone resulting from BP cloning between pDONR221 and *H. eriocheir* Hxk second PCR product.

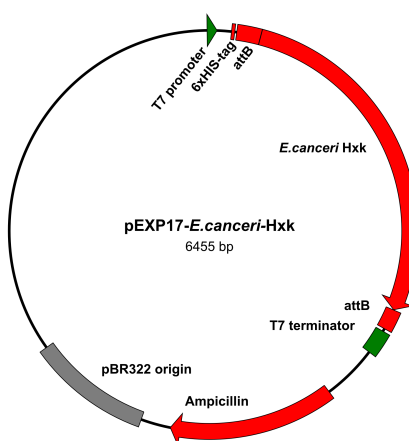


Figure 6: Schematic representation of the pEXP17-*Ent. canceri*-Hxk clone resulting from LR cloning between pENT221-*Ent. canceri*-Hxk and pDEST17.

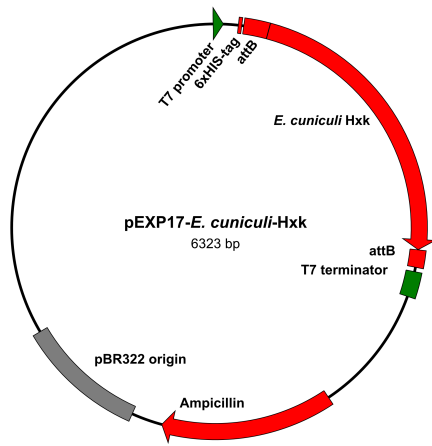


Figure 7: Schematic representation of the pEXP17-*Enc. cuniculi*-Hxk clone resulting from LR cloning between pENT221-*Enc. cuniculi*-Hxk and pDEST17.

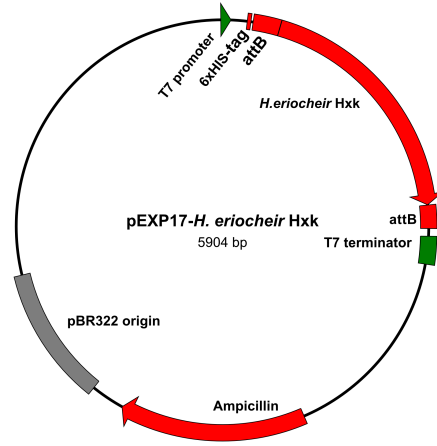


Figure 8: Schematic representation of the pEXP17-*H. eriocheir*-Hxk clone resulting from LR cloning between pENT221-*H. eriocheir*-Hxk and pDEST17

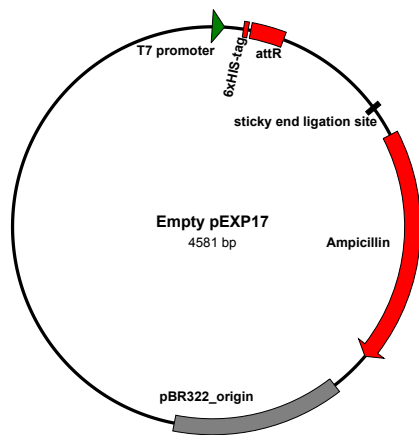


Figure 9: Schematic representation of the empty pEXP17 clone resulting from EcoRI restriction digestion.

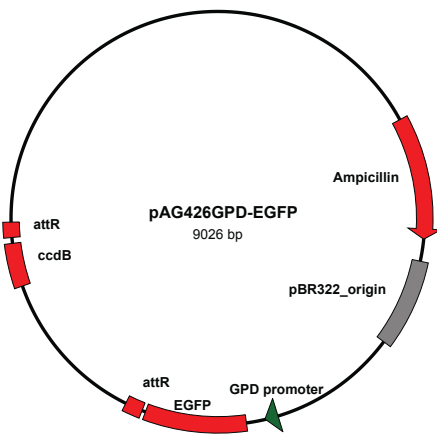


Figure 10: Schematic representation of the pAG426GPD-EGFP Gateway plasmid kindly donated by Graham Thomas, University of Exeter.

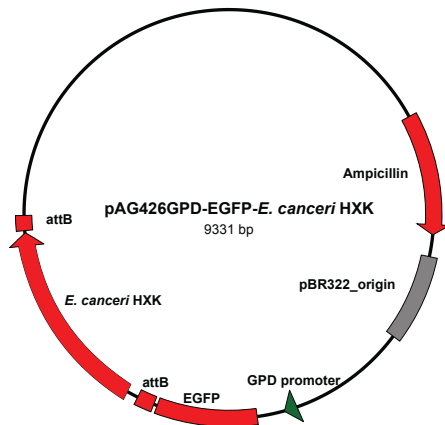


Figure 11: Schematic representation of the pAG426GPD-EGFP -*Ent. canceri*-Hxk clone resulting from LR cloning between pENT221-*Ent. canceri*-Hxk and pAG426GPD-EGFP

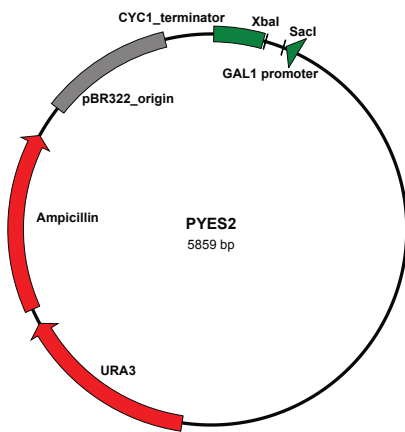


Figure 12: Schematic representation of the PYES2 expression vector (ThermoFisher Scientific, UK)

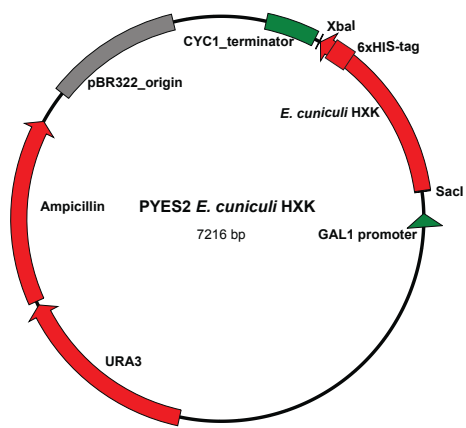


Figure 13: Schematic representation of the PYES2-*Enc. cuniculi*-Hxk clone resulting from stickyend ligation

Appendix 3: List of oligonucleotides used in this study

Table 1: Sequences of primers and their use in this study

Primer name	Primer sequence (5'-3')	Annealing temperature	Purpose
GWE_ca-p1F	ACAAAAAAGCAGGCTTCatgttctgccttgtttct a	67 °C	1 st Gateway PCR reaction: Amplification of <i>Ent. canceri</i> hexokinase without its PTPA domain from genomic DNA and adjoining part of the Gateway attB recombination site to its 5' terminal
GWE_ca-p1R	ACAAGAAAGCTGGGTcaccatgttcattacg		1 st Gateway PCR reaction: Amplification of <i>Ent. canceri</i> hexokinase without its PTPA domain from genomic DNA and adjoining part of the Gateway attB recombination site to its 3' terminal
GWE_ca1F	ACAAAAAAGCAGGCTTCatgactgactgcaaa attgg	67 °C	1 st Gateway PCR reaction: Amplification of <i>Ent. canceri</i> hexokinase from genomic DNA and adjoining part of the Gateway attB recombination site to its 5' terminal
GWE_ca1R	ACAAGAAAGCTGGGTcaccatgttcattacga attc		1 st Gateway PCR reaction: Amplification of <i>Ent. canceri</i> hexokinase from genomic DNA and adjoining part of the Gateway attB recombination site to its 3' terminal
GWp1F	ACAAAAAAGCAGGCTTCatgtttactatgcattta gc	66 °C	1 st Gateway PCR reaction: Amplification of <i>Enc. cucinuli</i> hexokinase from genomic DNA and adjoining part of the Gateway attB recombination site to its 5' terminal
GWp1R	ACAAGAAAGCTGGGTcaggagtctgatcgca aatga		1 st Gateway PCR reaction: Amplification of <i>Enc. cucinuli</i> hexokinase from genomic DNA and adjoining part of the Gateway attB recombination site to its 3' terminal
GWH_e1F	ACAAAAAAGCAGGCTTCatggaattaa agaact	42 °C	1 st Gateway PCR reaction: Amplification of <i>H. eriocheir</i> hexokinase from synthesised gene and adjoining part of the Gateway attB recombination site to its 5' terminal
GWH_e1R	ACAAGAAAGCTGGGTctatttaaattctataa att		1 st Gateway PCR reaction: Amplification of <i>H. eriocheir</i> hexokinase from synthesised gene and adjoining part of the Gateway attB recombination site to its 3' terminal
GWp2F	GGGACAAGTTTGTACAAAAAAGCAGGC TTC	64.4 °C	2 nd Gateway PCR reaction: Adjoining remaining length of BP recombination site to the 5' terminal of 1 st PCR product
GWp2R	GGGGACCACTTTGTACAAGAAAGCTGG GTC		2 nd Gateway PCR reaction: Adjoining remaining length of BP recombination site to the 3' terminal of 1 st PCR product
HA1Forward	GGTGGTTCTAGAATGTTTACTATGCATT T	60.4 °C	Primers to isolate hexokinase from <i>Enc. cucinuli</i> + adding poly-his tag to 3' end + adding XbaI site to 5' end
HA1Reverse	ATGATGATGATGATGAGGAGTCTGATC GGC		

HA2Forward/ HA1Forward	GGTGGTTCTAGAATGTTTACTATGCATT T	60.4 °C	Primers to extend histidine-tag + adding SacI restriction site to 3' end
HA2Reverse	GGTGGTGAGCTCATGATGATGATGATG ATG		
Primers for GATC sequencing			
M13-FP	TGTAAAACGACGGCCAGT	N/A	Sequencing insert in Gateway pENT221 clone forward primer
M13-RP	CAGGAAACAGCTATGACC	N/A	Sequencing insert in Gateway pENT221 clone reverse primer
T7	TAATACGACTCACTATAGGG	N/A	Sequencing insert in Gateway pEXP17 clone forward primer
GATC-primer- 1041334	CCGCTGAGCAATAACTAGC	N/A	Sequencing insert in Gateway pEXP17 clone reverse primer

Appendix 4: Bash scripts and partition files for 21-protein phylogenetic analysis

Maximum likelihood analysis on 21-protein concatenated alignment

Partition file

```
LGF, 986 = 1-323
LG, 987 = 324-690
LG, 992 = 691-899
LGF, 994 = 900-1060
LG, 1001 = 1061-1374
LG, 1005 = 1375-1461
LG, 1008 = 1462-1599
LGF, 1009 = 1600-1769
LGF, 1010 = 1770-1977
LGF, 1013 = 1978-2321
LGF, 1014 = 2322-2429
LGF, 1015 = 2430-3013
RTREVF, 1021 = 3014-3253
LG, 1023 = 3254-3406
LGF, 1024 = 3407-3637
LGF, 1026 = 3638-3934
LGF, 1027 = 3935-4062
LG, 1028 = 4063-4337
LGF, 1030 = 4338-4508
LG, 1039 = 4509-4725
LGF, 1043 = 4726-4950
```

The final analysis was performed with the following command:

```
raxmlHPC-PTHREADS-SSE3 -T 18 -f a -m PROTGAMMALGF -q
concatenated_multiplemodel_file.txt -# 100 -x 12345 -p 54321
-s input_concatenated_alignment_file -n output_file
```

Bayesian inference analysis on 21-protein concatenated alignment.

Dividing data into partitions

The command line version of MRBAYES (v3.6) (Ronquist et al. 2012) was employed. ./mb

Partition file

```
#NEXUS

begin mrbayes;
execute concatenated_alignment.nex;
charset spt = 1-323;
charset sec = 324-690;
charset pri = 691-899;
charset enp = 900-1060;
charset wrs = 1061-1374;
charset taf = 1375-1461;
charset tfa = 1462-1599;
charset vin = 1600-1769;
charset abd = 1770-1977;
charset secc = 1978-2321;
charset arh = 2322-2429;
charset sptt = 2430-3013;
charset brn = 3014-3253;
charset nob = 3254-3406;
charset caf = 3407-3637;
charset tfb = 3638-3934;
charset bos = 3935-4062;
charset npi = 4063-4337;
charset tma = 4338-4508;
charset tfaa = 4509-4725;
charset clp = 4726-4950;
partition favored = 21: spt, sec, pri, enp, wrs, taf, tfa,
vin, abd, secc, arh, sptt, brn, nob, caf, tfb, bos, NpI,
tma, tfaa, clp;
set partition = favored;
end;
```

Subsequently, this command file was used to load the nexus-formatted concatenated alignment by typing the following into terminal:

```
execute command_file.nex
```

Assigning evolutionary models to partititons

The priors across all partitions were unlinked by typing the following into terminal:

```
unlink statefreq=(all) revmat=(all) pinvar=(all)
```

and the rate prior was set to “variable” with the following command:

```
prset applyto=(all) ratepr=variable
```

The molecular evolutionary models estimated for each partition were applied with the following command:

```
lset          applyto=(partition#,partition#...)          prset  
aamodelpr=(estimated evolutionary model)
```

```
lset          applyto=(partition#,partition#...)          prset  
rates=gamma/invgamma
```

The sump command was used to assess the parameter values

Appendix 5: Bash scripts for plasma membrane transporter prediction

```
#!/bin/sh
WOLFPSORT="/home/dominic/tools/WoLFPSORT_package_v0.2/bin/r
unWolfPsortSummary" #where your WoLFPSORT script is found
ORG="fungi" #what organism is it? fungi, plant or animal
TMHMM="/mnt/Dominic/tmhmm-2.0c/bin/tmhmm"
SIGNALP="/mnt/Dominic/signalp-4.1/signalp"
TRANSPORTERBLASTDB="/mnt/Dominic/TransporterProteinDB/tcdb"
NCBINR="/mnt/Dominic/mydatabases/nr"
SGD="/mnt/Dominic/SGD_database/sgd.fa"
BAC="/mnt/Dominic/rickettsia_db/rickettsia.fasta"
for filename in ./merged5.fas.3 #please write the location
and name of you protein fasta file
do
echo "This is Dominic's Plasma Membrane transporter
Prediction Program"
echo "Using TMHMM to predict proteins with transmembrane
domains (TMD)-it takes 0.0007 mins per protein sequence"
"$TMHMM" "$filename" -short > "$filename".tmhmm #predicted
TM proteins
grep "PredHel=[1-9].*" < "$filename".tmhmm >
"$filename".tmhmm1 # remove proteins without transmembrane
domains
grep 'PredHel=1[[:space:]]' < "$filename".tmhmm1 >
"$filename".tmhmm2 # get proteins with single transmembrane
domain
sed 's/\t.*$//g' < "$filename".tmhmm2 > "$filename".tmhmm3
#get list of proteins with single transmembrane domain
grep -A 1 -f "$filename".tmhmm3 < "$filename" >
"$filename".tmhmm4 #get list of proteins with single
transmembrane domain in fasta
tr -d '-' < "$filename".tmhmm4 > "$filename".tmhmm5 #remove
"-" character from grep output file
```

```

sed '/^$/d' < "$filename".tmhmm5 > "$filename".tmhmm1.fasta
#Remove empty lines. fasta file of proteins with single
transmembrane domains
rm -f "$filename".tmhmm5 "$filename".tmhmm4
"$filename".tmhmm3 #remove useless files
echo "Using SIGNALP to find proteins that are predicted to
have a single TMD but also a Signal Peptide-They will be
removed"
"$SIGNALP" -t euk -f short "$filename".tmhmm1.fasta >
"$filename".tmhmm1.fasta.signalp #find out which of the
proteins with single transmembrane domains have a signal
petide
grep " Y " < "$filename".tmhmm1.fasta.signalp >
"$filename".tmhmm1.fasta.signalp1 #get proteins with single
TMD and signal petides
sed 's/#.*/#/'g < "$filename".tmhmm1.fasta.signalp1 >
"$filename".tmhmm1.fasta.signalp2 #get list of proteins
with single TMD and signal petides
grep -v -f "$filename".tmhmm1.fasta.signalp2 <
"$filename".tmhmm1 > "$filename".tmhmm1.true #remove
proteins with predicted single TMD and signal peptide from
initial list of predicted TM proteins
echo "All proteins with predicted TMD can be found in
"$filename".tmhmm1.true"
grep 'PredHel=[1-9].*' < "$filename".tmhmm1.true >
"$filename".tmhmm1.true.trans
sed 's/#.*/#/'g < "$filename".tmhmm1.true.trans >
"$filename".tmhmm1.true.trans1
grep -A 1 -f "$filename".tmhmm1.true.trans1 < "$filename" |
tr -d '-' | sed '/^$/d' >
"$filename".tmhmm1.true.trans1.fasta
rm -f "$filename".tmhmm1.true.trans1
echo "SEARCHING TCDB DATABASE FOR MATCHES"

```

```

blastall -p blastp -d "$TRANSPORTERBLASTDB" -i
"$filename".tmhmm1.true.trans1.fasta -e 1e-5 -m 8 -o
"$filename".tmhmm1.true.trans1.fasta.blastout.tcdb
echo "SEARCHING SGD DATABASE FOR MATCHES"
blastall -p blastp -d "$SGD" -i
"$filename".tmhmm1.true.trans1.fasta -e 1e-5 -m 8 -o
"$filename".tmhmm1.true.trans1.fasta.blastout.sgd
echo "SEARCHING BACTERIA DATABASE FOR MATCHES"
blastall -p blastp -d "$BAC" -i
"$filename".tmhmm1.true.trans1.fasta -e 1e-5 -m 8 -o
"$filename".tmhmm1.true.trans1.fasta.blastout.rick
echo "SORTING BLAST OUTPUT"
sort -buk1,1 <
"$filename".tmhmm1.true.trans1.fasta.blastout.tcdb >
"$filename".tmhmm1.true.trans1.fasta.blastout.tcdb.1
sort -buk1,1 <
"$filename".tmhmm1.true.trans1.fasta.blastout.sgd >
"$filename".tmhmm1.true.trans1.fasta.blastout.sgd.1
sort -buk1,1 <
"$filename".tmhmm1.true.trans1.fasta.blastout.rick >
"$filename".tmhmm1.true.trans1.fasta.blastout.rick.1
awk '!a[$1]++' "$filename".tmhmm.true.trans.fasta.blastout
> "$filename".tmhmm.true.trans1.fasta.blastout2
done
echo "CONVERTING BLAST RESULTS INTO PROTEIN FAMILIES"
for filename in /*.trans1.fasta.blastout.*1
do
cut -f 1,2 < "$filename" > "$filename".2
done
echo ">FORMATTING RICKETTSSIA BLAST OUTPUT"
for filename in
/*.tmhmm1.true.trans1.fasta.blastout.rick.1.2
do
sed 's/#\t..|/#\t/'g < "$filename" > "$filename".3
sed 's/|.*$// 'g < "$filename".3 > "$filename".3.4

```

```

sed -f ./rick.txt < "$filename".3.4 > "$filename".3.4.5
cut -f 1 < "$filename".3.4.5 > test.rick.txt
done
echo ">FORMATTING SGD BLAST OUTPUT"
for filename in
./*.tmhmm1.true.trans1.fasta.blastout.sgd.1.2
do
sed -f ./sgd.txt < "$filename" > "$filename".3
cut -f 1 < "$filename".3 > test.sgd.txt
done
echo ">FORMATTING TCDB BLAST OUTPUT"
for filename in
./*.tmhmm1.true.trans1.fasta.blastout.tcdb.1.2
do
sed 's/[0-9]*.[0-9]*$/#/'g < "$filename" > "$filename".3
sed -f ./tcdb.txt < "$filename".3 > "$filename".3.4
cut -f 1 < "$filename".3.4 > test.tcdb.txt
done
echo "MERGING BLAST OUTPUTS"
grep -v -f test.sgd.txt < test.rick.txt > test.txt
cat test.sgd.txt test.txt > merged1
grep -v -f merged1 < test.tcdb.txt > merged2
cat merged1 merged2 | sort > merged3
cp merged3 test.txt
echo ">>CREATING sed FILES"
for filename in ./*.tmhmm1.true.trans1.fasta
do
grep -v "[0-9]$" < "$filename".blastout.rick.1.2.3.4.5 >
"$filename".blastout.rick.1.2.3.4.5.6
sed 's/^/s\\/'g < "$filename".blastout.rick.1.2.3.4.5.6 |
sed 's/#\t#\//'g | sed 's/$/\g/'g >
"$filename".blastout.rick.1.2.3.4.5.6.7
grep -v "[0-9]$" < "$filename".blastout.sgd.1.2.3 >
"$filename".blastout.sgd.1.2.3.4

```

```

sed 's/^/s\\/'g < "$filename".blastout.sgd.1.2.3.4 | sed
's/#\t/#\\/'g | sed 's/$/\g/'g >
"$filename".blastout.sgd.1.2.3.4.5
grep -v "#$" < "$filename".blastout.tcdb.1.2.3.4 >
"$filename".blastout.tcdb.1.2.3.4.5
sed 's/^/s\\/'g < "$filename".blastout.tcdb.1.2.3.4.5 | sed
's/#\t/#\\/'g | sed 's/$/\g/'g >
"$filename".blastout.tcdb.1.2.3.4.5.6
sed -f "$filename".blastout.sgd.1.2.3.4.5 < test.txt >
test1.txt
sed -f "$filename".blastout.rick.1.2.3.4.5.6.7 < test1.txt
> test2.txt
sed -f "$filename".blastout.tcdb.1.2.3.4.5.6 < test2.txt >
test3.txt
sed 's/.*$$/unknown/'g < test3.txt > test4.txt
paste merged3 test4.txt > test5.txt
sed 's/[0123456789].*#/'g < test5.txt > test6.txt
awk '{count[$1]++}END{for(j in count) print j, "("count[j]"
proteins)"}' FS=: test6.txt > test7.txt
sort test7.txt > DoPMet.predTMT.pfam
echo "EXTRACTING FASTA OF PREDICTED TM TRANSPORTER PROTEINS
(TMT)"
grep -A 1 -f merged3 < "$filename" | tr -d '-' > tmp
sed 's/#$/_/'g < tmp > tmp1
sed '/^$/d' < tmp1 > "$filename".predTMT.fa
echo "WOLFPSORT IS PREDICTING INTRACELLULAR LOCALIZATION"
"$WOLFPSORT" "$ORG" < "$filename".predTMT.fa >
"$filename".wolfpsort
grep '_ plas.*$' < "$filename".wolfpsort |sort|uniq >
"$filename".predTMT.fa.plasMem
sed 's/_ plas .*/#/'g < "$filename".predTMT.fa.plasMem >
tmp
grep -f tmp < test5.txt >
"$filename".predTMT.fa.plasMem.pfam

```

```
sed 's/[0123456789].*#// 'g <
"$filename".predTMT.fa.plasMem.pfam > tmp
awk '{count[$1]++}END{for(j in count) print j, "("count[j]"
proteins)"}' FS=: tmp > tmp1
sort tmp1 > DoPMet.output.plasMemProt.Pfam.count
grep -v -f trimming_file.txt <
DoPMet.output.plasMemProt.Pfam.count >
DoPMet.output.plasMemProt.Pfam.count.trim
rm -f tmp
rm -f tmp1
echo "Dominic's Membrane Transporter protein prediction
program has finished successfully!"
done
```

Appendix 6: Secretome predictions for the Microsporidia

Ortholog clusters of microsporidian secreted proteins

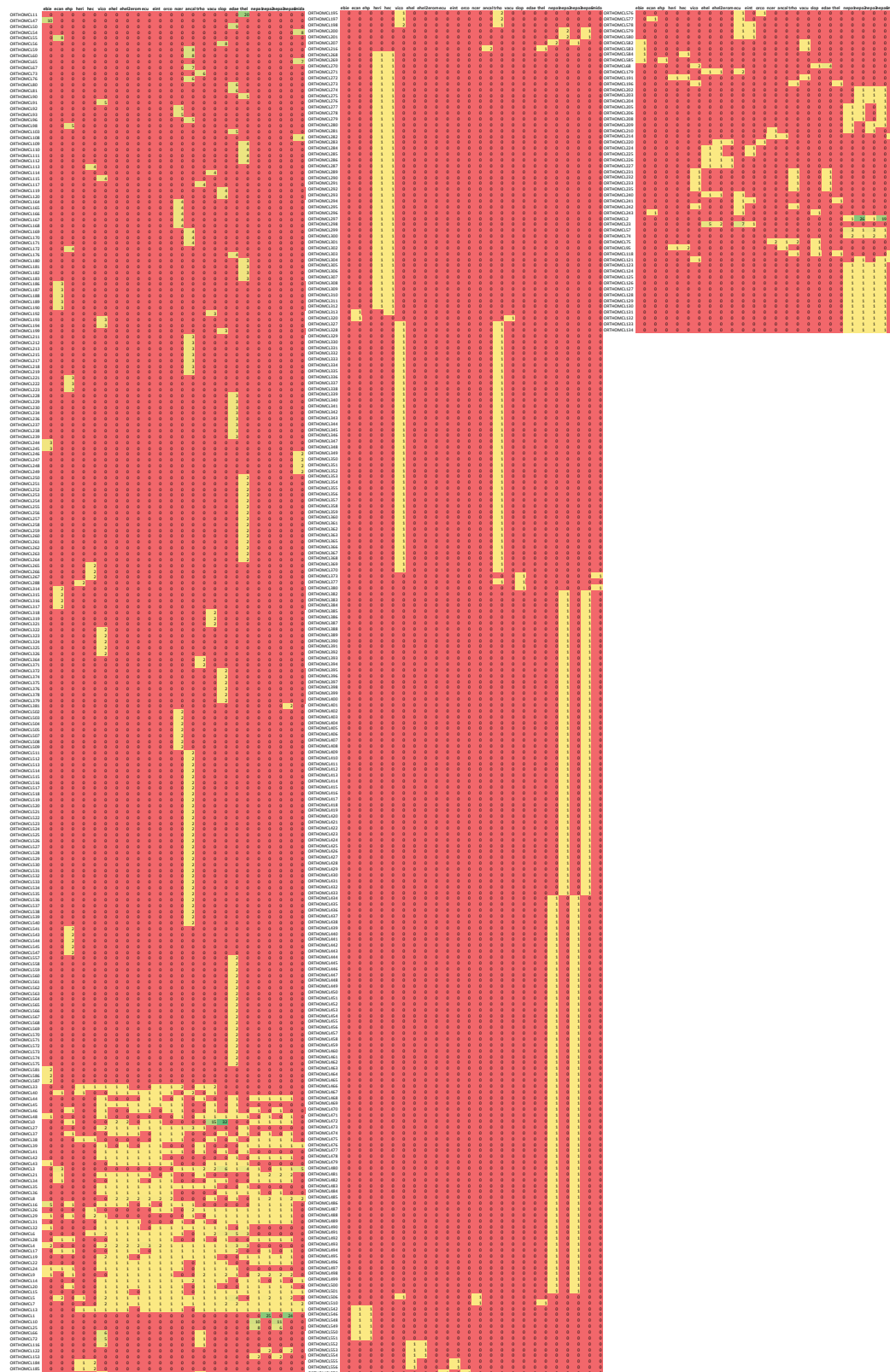


Figure 1: Heat map of ortholog clusters of predicted secreted proteins from 23 microsporidian species including species sequenced in this study. This heat map demonstrates that none of the secreted proteins predicted here are ubiquitous across the microsporidian phylum

Table 1: Statistics for the predicted secretome of 23microsporidian species

	<i>E. bienersi</i>	<i>Ent. canceri</i>	<i>E. hepatopernai</i>	<i>H. eriocheir</i>	<i>H. eriocheir canceri</i>	<i>V. comeae</i>	<i>Enc. hellen Swiss</i>	<i>Enc. hellen ATCC</i>	<i>Enc. romaleae</i>	<i>Enc. cuniculi</i>	<i>Enc. intestinalis</i>	<i>O. colligata</i>	<i>N. ceranae</i>	<i>A. algerae</i>	<i>T. hominis</i>	<i>Vav. culicis</i>	<i>S. lophii</i>	<i>Edh. aedis</i>	<i>Thelephania spp.</i>	<i>Nem. Parisii ERTM1</i>	<i>Nematocida sp. ERTM2</i>	<i>Nem. Parisii ERTM3</i>	<i>Nematocida sp. ERTM6</i>	<i>M. daphniae</i>
orthologous copies (copies of gene present in more than one taxa)	20	24	23	56	62	117	80	75	56	83	76	57	46	32	100	55	81	44	43	172	190	170	182	20
unique copies (single copies present in only one taxa)	41	59	98	24	24	82	8	0	5	13	7	11	67	91	75	54	71	221	61	11	12	11	11	113
unique paralogous copies (multiple copies of same gene present in only 1 taxa)	22	31	28	2	10	25	0	0	0	0	0	0	44	127	14	13	31	92	83	0	0	0	2	27
total	83	114	149	82	96	224	88	75	61	96	83	68	157	250	189	122	183	357	187	183	202	181	195	160
total number of proteins encoded by genome	3806	2179	2540	2716	3058	2340	2006	1864	1883	2029	2011	1879	2678	3661	3253	2880	2596	4281	6226	2724	2831	2788	2484	3428
proportion of proteins that are secreted	0.02	0.05	0.05	0.03	0.03	0.09	0.04	0.04	0.03	0.04	0.04	0.03	0.05	0.06	0.06	0.04	0.07	0.08	0.03	0.07	0.07	0.07	0.08	0.05

Appendix 7: Secretome prediction bash scripts

```
#!/bin/sh
#all.fasta.tmhmm.0_1.signalp.1.list.orthocluster.1.onlySignalp
alProts this is the orthomcl output file containing IDs of
proteins predicted to have signal peptides
SIGNALP="/home/dominic/Documents/signalp-4.1/signalp"
TMHMM="/home/dominic/Documents/tmhmm-2.0c/bin/tmhmm"
for filename in ./all.fasta #please write the location and
name of your protein fasta file
do
echo "This is Dominic's secretome Prediction Program-DoSe"
echo "Using TMHMM to predict proteins with transmembrane
domains (TMD)-it takes 0.0007 mins per protein sequence"
"$TMHMM" "$filename" -short > "$filename".tmhmm #predicted
TM proteins
grep 'PredHel=[0-1]\t.*$' < "$filename".tmhmm >
"$filename".tmhmm.0_1 # get proteins with 0 or 1
transmembrane domain
echo ">>>sed<<<"
sed 's/#.*$/#/'g < "$filename".tmhmm.0_1 > delete_this0 #
create list of proteins with 0 or 1 transmembrane domain
echo ">>>sort<<<"
sort < delete_this0 |uniq > "$filename".tmhmm.0_1.list
echo ">>>getting FASTA files please be patient<<<"
fgrep -A 1 -f "$filename".tmhmm.0_1.list < "$filename" >
delete_this1 # get proteins with 0 or 1 transmembrane
domain in fasta
echo ">>>editing FASTA files<<<"
tr -d '-' < delete_this1 > delete_this2 #remove "-"
character from grep output file
sed '/^$/d' < delete_this2 > "$filename".tmhmm.0_1.fa
#Remove empty lines. fasta file of proteins with single
transmembrane domains
echo "Using SIGNALP to find proteins with a Signal Peptide"
```

```

"$SIGNALP" -t euk -f short "$filename".tmhmm.0_1.fa >
"$filename".tmhmm.0_1.fa.signalp #find out which of the
proteins have a signal petide
grep " Y " < "$filename".tmhmm.0_1.fa.signalp >
delete_this3.1 #get proteins with signal petides
sed 's/#.*#/'g < delete_this3.1 > delete_this3.2
sort < delete_this3.2 |uniq >
"$filename".tmhmm.0_1.fa.signalp.only #get list of proteins
with signal peptides only
grep " N " < "$filename".tmhmm.0_1.fa.signalp >
delete_this3 #get proteins with NO signal petides
sed 's/#.*#/'g < delete_this3 >
"$filename".tmhmm.0_1.fa.Nosignalp.only #get list of
proteins with NOsignal peptides only
echo "creating SED file to remove proteins without signal
peptides from orthoMCL output"
sed 's/^/s\\/'g < "$filename".tmhmm.0_1.fa.Nosignalp.only
|sed 's/$/\\/'g >
"$filename".tmhmm.0_1.fa.Nosignalp.only.sed
echo "removing proteins with No signal peptides from
OrthoMCL output"
sed -f "$filename".tmhmm.0_1.fa.Nosignalp.only.sed <
all_orthomcl.out > delete_this4
echo "getting proteins with more than 1 transmembrane
domain"
grep 'PredHel=[2-9]\\t' < "$filename".tmhmm >
"$filename".tmhmm.2_more
echo "createing list of proteins with more than 1
transmembrane domain"
sed 's/#.*$#/'g < "$filename".tmhmm.2_more >
"$filename".tmhmm.2_more.list # create list of proteins
with more than 1 transmembrane domain
echo ">>>creating sed file<<<"
sed 's/^/s\\/'g < "$filename".tmhmm.2_more.list |sed
's/$/\\/'g > "$filename".tmhmm.2_more.list.sed

```

```

echo "removing proteins with more than 1 transmembrane
domain"
sed -f "$filename".tmhmm.2_more.list.sed < delete_this4 >
"$filename".tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP
done
echo "THIS PART IS COUNTING THE NUMBER OF ORTHOLOGS IN EACH
CLUSTER FOR EACH SPECIES"
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/EBI_/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP > 1.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/ECANGB/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP > 2.txt
awk 'BEGIN{print "count", "lineNum"}{print gsub(/EHP/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP > 3.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/HERIO/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP > 4.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/A0H76/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP > 5.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/VICG/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP > 6.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/EHEL/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP > 7.txt
awk 'BEGIN{print "count", "lineNum"}{print gsub(/KMI/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP > 8.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/EROM/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP > 9.txt
awk 'BEGIN{print "count", "lineNum"}{print gsub(/ECU/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
10.txt

```

```

awk 'BEGIN{print "count", "lineNum"}{print
gsub(/Eint/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
11.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/M896/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
12.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/NCER/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
13.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/H311/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
14.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/THOM_/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
15.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/VCUG/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
16.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/SLOPH/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
17.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/EDEG/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
18.txt
awk 'BEGIN{print "count", "lineNum"}{print gsub(/the1[0-
9]/, "")}'

```

```

all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
19.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/NEPG/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
20.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/NERG/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
21.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/NEQG/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
22.txt
awk 'BEGIN{print "count", "lineNum"}{print
gsub(/NESG/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
23.txt
awk 'BEGIN{print "count", "lineNum"}{print gsub(/tr_/, "")}'
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP >
24.txt
paste 1.txt 2.txt 3.txt 4.txt 5.txt 6.txt 7.txt 8.txt 9.txt
10.txt 11.txt 12.txt 13.txt 14.txt 15.txt 16.txt 17.txt
18.txt 19.txt 20.txt 21.txt 22.txt 23.txt 24.txt > tmp
sed '1d' tmp > merged.txt
cut -f 1
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP > tmp1
paste tmp1 merged.txt >
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP.txt
#THIS FILE CAN BE OPENED IN EXCEL. IT CONTAINS COUNTS OF
ORTHOLOGOUS SECRETED PROTEINS FOR EACH SPECIES.
cut -f 2 <
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP > tmp2
tr ' ' '\n' < tmp2 > tmp3
sed 's/(.*//g' < tmp3 > tmp4

```

```
sort tmp4 | sed '/^$/d' >
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP.IDs
fgrep -v -f
all.fasta.tmhmm.0_1.fa.signalp.orthoMCl.onlySignalP.IDs <
all.fasta.tmhmm.0_1.fa.signalp.only >
all.fasta.tmhmm.0_1.signalp.1.list.Noorthocluster.1.onlySig
nalpProts.IDs
cut -c-4
all.fasta.tmhmm.0_1.signalp.1.list.Noorthocluster.1.onlySig
nalpProts.IDs > tmp6
awk '{count[$1]++}END{for(j in count) print j,"("count[j]"
proteins)"}' FS=: tmp6 >
all.fasta.tmhmm.0_1.signalp.1.list.Noorthocluster.1.onlySig
nalpProts.txt #THIS FILE CAN BE OPENED IN EXCEL. IT
CONTAINS COUNTS OF NON-ORTHOLOGOUS SECRETED PROTEINS FOR
EACH SPECIES.
```

Appendix 8: Matrix-assisted laser desorption/ionization Mass Spectrometry (MALDI-MS) analysis

Table 1: Summary of proteins detected by MALDI-MS analysis for the 50 kDa gel band excised during recombinant *Enc.uniculi* protein expression in *S. cerevisiae*.

Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide	Protein MW (Da)	Species	Protein Name
1	11	10	143.69	32.5	4.82E+05	50431.1	YEAST	RecName: Full=Elongation factor 1-alpha; Short=EF-1-alpha; AltName: Full=Eukaryotic elongation factor 1A; Short=eEF1A; AltName: Full=Translation elongation factor 1A
1	11	10	143.69	32.5	4.82E+05	50431.1	Saccharomyces cerevisiae	Tef1p: Elongation factor 1-alpha
1	11	10	143.69	32.5	4.82E+05	50431.1	Saccharomyces cerevisiae	elongation factor 1-alpha
1	11	10	143.69	32.5	4.82E+05	50431.1	Saccharomyces cerevisiae	EF-1-alpha
1	11	10	143.69	32.5	4.82E+05	50431.1	Saccharomyces cerevisiae	unnamed protein product
1	11	10	143.69	32.5	4.82E+05	50431.1	Saccharomyces cerevisiae	elongation factor EF-1-alpha
1	10	9	128.61	37.2	4.89E+05	41566.8	Saccharomyces cerevisiae	translation elongation factor 1-alpha, partial
1	10	9	128.61	37.2	4.89E+05	41566.8	Saccharomyces cerevisiae	translation elongation factor 1-alpha
1	10	9	128.61	37.2	4.89E+05	41566.8	Saccharomyces cerevisiae	translation elongation factor 1-alpha
1	10	9	128.61	37.2	4.89E+05	41566.8	Saccharomyces cerevisiae	translation elongation factor 1-alpha
1	10	9	128.61	38.1	4.89E+05	40593.7	Saccharomyces cerevisiae	translation elongation factor 1-alpha
1	9	8	119.22	29.5	5.38E+05	41574.9	Saccharomyces cerevisiae	translation elongation factor 1-alpha
1	7	7	92.13	40.3	5.16E+05	32844.9	Saccharomyces cerevisiae	translation elongation factor 1-alpha, partial
1	5	5	71.64	13.7	4.63E+05	50267	ASHGO	RecName: Full=Elongation factor 1-alpha; Short=EF-1-alpha
1	4	4	61.22	12.1	4.25E+05	50539.3	YARLI	RecName: Full=Elongation factor 1-alpha; Short=EF-1-alpha
1	4	4	61.22	12.1	4.25E+05	50525.3	Yarrowia lipolytica	translation elongation factor 1-alpha
1	3	3	51.83	5.8	5.49E+05	50467.3	Candida albicans SC5314	probable translation elongation factor EF-1 alpha
1	3	3	51.83	5.8	5.49E+05	50467.3	Candida albicans SC5314	probable translation elongation factor EF-1 alpha
1	3	3	51.83	5.8	5.49E+05	50467.3	Candida albicans SC5314	probable translation elongation factor EF-1 alpha
1	3	3	51.83	5.8	5.49E+05	50467.3	Candida albicans SC5314	probable translation elongation factor EF-1 alpha
1	3	3	51.83	5.8	5.49E+05	50455.2	dubliniensis CD36	ef-1-alpha, putative
1	3	3	51.83	5.8	5.49E+05	50455.2	dubliniensis CD36	ef-1-alpha, putative; elongation factor 1-alpha, putative; eukaryotic elongation factor 1a, putative; translation elongation factor 1a, putative
1	3	3	51.83	5.8	5.49E+05	50427.2	dubliniensis CD36	translation elongation factor 1-alpha, putative
1	3	3	51.83	5.8	5.49E+05	50427.2	dubliniensis CD36	translation elongation factor 1-alpha, putative
1	2	2	33.3	4.3	6.63E+05	50441.2	Candida albicans SC5314	probable translation elongation factor EF-1 alpha
1	2	2	33.3	4.3	6.63E+05	50441.2	Candida albicans SC5314	probable translation elongation factor EF-1 alpha
1	2	2	33.3	4.3	6.63E+05	50441.2	Candida albicans SC5314	probable translation elongation factor EF-1 alpha
1	2	2	33.3	4.3	6.63E+05	50441.2	Candida albicans SC5314	probable translation elongation factor EF-1 alpha
2	10	10	131.88	31.4	2.19E+05	30949.9	<i>Escherichia coli</i> str. K-12 substr. W3110	D-ribose transporter subunit
2	10	10	131.88	31.4	2.19E+05	30949.9	<i>Escherichia coli</i> str. K12 substr. W3110	D-ribose transporter subunit
3	3	3	57.06	17	5.22E+06	24159.8	UNREADABLE_PD	b 1LDT T Chain T
3	2	2	37.49	12.5	7.36E+06	24158.8	UNREADABLE_PD	b 1AN1 E Chain E
4	3	3	45.56	1.5	2.89E+05	203669	Saccharomyces cerevisiae	Yhr214c-bp
4	3	3	45.56	1.5	2.89E+05	199457	YEASX	RecName: Full=Transposon Full=Pol-p63; AltName: Full=p60
4	3	3	45.56	1.5	2.89E+05	199421	Saccharomyces cerevisiae	Yer160cp
4	3	3	45.56	1.5	2.89E+05	199357	Saccharomyces cerevisiae	transposon Ty putative with frame shift at 5704-5706
4	3	3	45.56	6.1	2.89E+05	49237.7	YEASX	RecName: Full=Transposon TyH3 Gag polyprotein;

4	3	3	45.56	6.1	2.89E+05	49220.7	YEAST	RecName: Full=Transposon Ty1-LR3 Gag polyprotein;
5	3	3	30.7	28.5	6.16E+04	13947.6	<i>Escherichia coli</i> str. K-12 substr. W3110	aspartate 1-decarboxylase
5	3	3	30.7	28.5	6.16E+04	13947.6	<i>Escherichia coli</i> str. K12 substr. W3110	aspartate 1-decarboxylase
6	1	1	23.06	33.3	3.04E+05	9225.4	<i>Escherichia coli</i> str. K-12 substr. W3110	HU, DNA-binding transcriptional regulator subunit beta
6	1	1	23.06	33.3	3.04E+05	9225.4	<i>Escherichia coli</i> str. K12 substr. W3110	HU, DNA-binding transcriptional regulator, beta subunit
7	2	2	22.54	3.3	4.36E+04	53420.7	<i>Saccharomyces cerevisiae</i>	GABA aminotransferase
8	2	2	20.59	5	5.36E+04	54780.7	<i>Saccharomyces arboricola</i> H-6	atp2p
8	2	2	20.59	5.4	5.36E+04	51125.4	UNREADABLE_PD B	b 2WPD D Chain D,
8	2	2	20.59	5.4	5.36E+04	51125.4	UNREADABLE_PD B	b 2WPD E Chain E,
8	2	2	20.59	5.4	5.36E+04	51125.4	UNREADABLE_PD B	b 2WPD F Chain F, b 3ZRY D Chain D, Rotor Architecture
8	2	2	20.59	5.4	5.36E+04	51125.4	UNREADABLE_PD B	In The F(1)-C(10)-Ring Complex Of The Yeast F-Atp Synthase
8	2	2	20.59	5.4	5.36E+04	51125.4	UNREADABLE_PD B	b 3ZRY E Chain E, Rotor Architecture
8	2	2	20.59	5.4	5.36E+04	51125.4	UNREADABLE_PD B	In The F(1)-C(10)-Ring Complex Of The Yeast F-Atp Synthase
8	2	2	20.59	5.4	5.36E+04	51125.4	UNREADABLE_PD B	b 3ZRY F Chain F, Rotor Architecture
8	2	2	20.59	5.4	5.36E+04	51125.4	UNREADABLE_PD B	In The F(1)-C(10)-Ring Complex Of The Yeast F-Atp Synthase
9	1	1	15.35	11	3.12E+04	11920.5	<i>Escherichia coli</i> str. K-12 substr. W3110	thioredoxin 1
9	1	1	15.35	11.1	3.12E+04	11920.5	UNREADABLE_PD B	b 1TXX A Chain A, Active-Site Variant Of E.Coli Thioredoxin
9	1	1	15.35	11	3.12E+04	11920.5	<i>Escherichia coli</i> str. K12 substr. W3110	thioredoxin 1
10	1	1	14.55	15.3	9.36E+04	15596.4	<i>Escherichia coli</i> str. K12 substr. W3110	global DNA-binding transcriptional dual regulator H-NS
10	1	1	14.55	15.3	9.36E+04	15596.4	<i>Escherichia coli</i> str. K-12 substr. W3110	global DNA-binding transcriptional dual regulator H-NS
11	1	1	13.47	3	4.58E+04	33019.3	<i>Fomitiporia mediterranea</i> MF3/22	CRAL/TRIO domain-containing protein
12	1	1	13.21	1.8	1.41E+05	49510.8	<i>Saccharomyces arboricola</i> H-6	gpd2p
12	1	1	13.21	2	1.41E+05	42732.9	<i>Saccharomyces cerevisiae</i>	glycerol-3-phosphate dehydrogenase (NAD+)
13	1	1	13.12	2	4.16E+04	43607.6	<i>Saccharomyces arboricola</i> H-6	gpd1p
13	1	1	13.12	2	4.16E+04	43438.6	UNREADABLE_PD B	b 4FGW A Chain A, Structure Of Glycerol-3-phosphate Dehydrogenase, Gpd1, From <i>Saccharomyces Cerevisiae</i>
13	1	1	13.12	2	4.16E+04	43438.6	UNREADABLE_PD B	b 4FGW B Chain B, Structure Of Glycerol-3-phosphate Dehydrogenase, Gpd1, From <i>Saccharomyces Cerevisiae</i>
13	1	1	13.12	2	4.16E+04	43438.6	YEAST	RecName: Full=Glycerol-3-phosphate dehydrogenase [NAD(+)] 1
14	1	1	11.85	1.4	5.94E+04	73305.7	<i>Candida dubliniensis</i> CD36	polyamine transporter, putative
14	1	1	11.85	1.4	5.94E+04	73305.7	<i>Candida dubliniensis</i> CD36	polyamine transporter, putative
15	1	1	11.73	1.6	9.66E+04	56633.7	<i>Gallus gallus</i>	asparagine-linked glycosylation protein 11 homolog
16	1	1	11.71	1.9	4.96E+04	50283.9	YEAST	RecName: Full=Cytochrome b-c1 complex subunit 1, mitochondrial; AltName: Full=Complex III subunit 1; AltName: Full=Core protein I; AltName: Full=Ubiquinol-cytochrome-c
16	1	1	11.71	2	4.96E+04	47364.6	UNREADABLE_PD B	reductase complex core protein 1; Flags: Precursor
16	1	1	11.71	2	4.96E+04	47364.6	UNREADABLE_PD B	b 1KYO A Chain A, Yeast Cytochrome Bc1 Complex With Bound Substrate Cytochrome C
16	1	1	11.71	2	4.96E+04	47364.6	UNREADABLE_PD B	b 1KYO L Chain L, Yeast Cytochrome Bc1 Complex With Bound Substrate Cytochrome C
Totals	252	241						

Table 2: Summary of proteins detected by MALDI-MS analysis for the 25 kDa gel band excised during recombinant *Enc. cuniculi* protein expression in *S. cerevisiae*.

Group (#)	Spectra (#)	Distinct Peptides (#)	Distinct Summed MS/MS Search Score	% AA Coverage	Mean Peptide Intensity	Protein MW (Da)	Species	Protein Name
1	3	3	62.27	17	3.46E+06	24159.8	UNREADABLE_PDB	b 1LDT T Chain T, Complex Of Leech-Derived Trypsin Inhibitor With Porcine Trypsin
1	2	2	44.35	12.5	4.80E+06	24158.8	UNREADABLE_PDB	b 1AN1 E Chain E, Leech-Derived Trypsin Inhibitor TRYPSIN COMPLEX
2	4	3	51.41	20	8.50E+04	22603.4	<i>Saccharomyces cerevisiae</i>	Rps8bp: Ribosome protein, small subunit
2	3	2	31.78	14	7.89E+04	22590.3	<i>Saccharomyces arboricola</i> H-6	rps8bp

2	3	2	31.78	14	7.89E+04	Saccharomyces arboricola H-6	22590.3	rps8ap
2	1	1	19.28	7.2	1.54E+05	Candida dubliniensis CD36	22809.5	ribosomal protein, small subunit, putative
2	1	1	19.28	7.2	1.54E+05	Candida dubliniensis CD36	22809.5	40S ribosomal protein S8
2	1	1	19.28	7.2	1.54E+05	Candida albicans SC5314	22809.5	likely cytosolic ribosomal protein S8
2	1	1	19.28	7.2	1.54E+05	Candida albicans SC5314	22809.5	likely cytosolic ribosomal protein S8
2	1	1	19.28	7.2	1.54E+05	Candida albicans SC5314	22809.5	likely cytosolic ribosomal protein S8
2	1	1	19.28	7.2	1.54E+05	Candida albicans SC5314	22809.5	likely cytosolic ribosomal protein S8
3	2	2	37.13	10.4	7.09E+04	<i>Escherichia coli</i> str. K-12 substr. W3110	30949.9	D-ribose transporter subunit
3	2	2	37.13	10.4	7.09E+04	<i>Escherichia coli</i> str. K12 substr. W3110	30949.9	D-ribose transporter subunit
4	2	2	32.83	8.3	1.41E+06	Saccharomyces arboricola H-6	25011	pnc1p
5	3	2	21.06	20.1	3.33E+04	Saccharomyces cerevisiae	12869.3	unnamed protein product
6	1	1	20.15	10.4	4.59E+04	bj3U5E S Chain S, The Structure Of The Eukaryotic Ribosome At 3.0 A Resolution. This Entry Contains Proteins Of The 60s Subunit, Ribosome A	20436.4	UNREADABLE_PDB
6	1	1	20.15	10.4	4.59E+04	bj3U5I S Chain S, The Structure Of The Eukaryotic Ribosome At 3.0 A Resolution. This Entry Contains Proteins Of The 60s Subunit, Ribosome B	20436.4	UNREADABLE_PDB
6	1	1	20.15	10.4	4.59E+04	bj4B6A S Chain S, Cryo-Em Structure Of The 60s Ribosomal Subunit In Complex With Arx1 And Rei1	20436.4	UNREADABLE_PDB
6	1	1	20.15	10.4	4.59E+04	bj4BYN S Chain S, Cryo-em Reconstruction Of The 80s-eif5b-met-itrnamet Eukaryotic Translation Initiation Complex	20436.4	UNREADABLE_PDB
6	1	1	20.15	10.4	4.59E+04	bj4BYU S Chain S, Cryo-em Reconstruction Of The 80s-eif5b-met-itrnamet Eukaryotic Translation Initiation Complex	20436.4	UNREADABLE_PDB
7	2	2	19.57	10.2	1.65E+04	Saccharomyces cerevisiae	20619.8	ribosomal protein L18
7	2	2	19.57	10.2	1.65E+04	YEAST	20619.8	RecName: Full=60S ribosomal protein L18-A; AltName: Full=RP28
8	2	2	17.49	6.5	6.56E+04	Saccharomyces arboricola H-6	24243.7	cdc33p
9	2	1	16.68	4.4	3.32E+05	Candida dubliniensis CD36	23037.3	60S ribosomal protein L13, putative
9	2	1	16.68	4.4	3.32E+05	Candida dubliniensis CD36	23037.3	60S ribosomal protein L13
10	2	1	16.02	6.6	2.62E+04	Saccharomyces arboricola H-6	25081.2	rps5p
11	1	1	12.7	3.4	2.62E+04	Saccharomyces cerevisiae YJM789	24478.8	ribosomal protein L15A
11	1	1	12.7	3.4	2.62E+04	Saccharomyces cerevisiae S288c	24478.8	TPA: ribosomal 60S subunit protein L15A
11	1	1	12.7	3.4	2.62E+04	Saccharomyces cerevisiae	24478.8	ribosomal protein YL10
11	1	1	12.7	3.4	2.62E+04	Saccharomyces cerevisiae S288c	24478.8	ribosomal 60S subunit protein L15A
11	1	1	12.7	3.4	2.62E+04	Saccharomyces cerevisiae Kyokai no. 7	24477.8	K7_Rpl15ap
11	1	1	12.7	3.4	2.62E+04	Saccharomyces cerevisiae	24477.8	CEN.PK113-7D Rpl15bp
11	1	1	12.7	3.4	2.62E+04	Saccharomyces arboricola H-6	24477.8	rpl15ap
11	1	1	12.7	3.4	2.62E+04	Candida dubliniensis CD36	24385.7	ribosomal protein, large subunit, putative
11	1	1	12.7	3.4	2.62E+04	Candida dubliniensis CD36	24385.7	60S ribosomal protein L15
11	1	1	12.7	3.4	2.62E+04	Candida albicans SC5314	24385.7	likely cytosolic ribosomal protein L15
11	1	1	12.7	3.4	2.62E+04	Candida albicans SC5314	24385.7	likely cytosolic ribosomal protein L15

11	1	1	12.7	3.4	2.62E+04	24385.7	Candida albicans SC5314	likely cytosolic ribosomal protein L15
11	1	1	12.7	3.4	2.62E+04	24385.7	Candida albicans SC5314	likely cytosolic ribosomal protein L15
11	1	1	12.7	3.4	2.62E+04	24378.7	Saccharomyces arboricola H-6	rpl15bp
12	1	1	11.99	3.9	5.85E+04	23145.8	Aspergillus niger	unnamed protein product
12	1	1	11.99	3.9	5.85E+04	23145.8	Aspergillus niger ATCC 1015	rpl16 ribosomal protein
12	1	1	11.99	4	5.85E+04	22704.3	Candida orthopsilosis	Rpl16a protein
12	1	1	11.99	4	5.85E+04	22704.3	Candida orthopsilosis Co 90-125	Rpl16a protein
12	1	1	11.99	4	5.85E+04	22692.3	Candida dubliniensis CD36	60S ribosomal protein L16, putative
12	1	1	11.99	4	5.85E+04	22692.3	Candida dubliniensis CD36	60S ribosomal protein L16
12	1	1	11.99	4	5.85E+04	22692.3	Candida albicans SC5314	likely cytosolic ribosomal protein L16
12	1	1	11.99	4	5.85E+04	22692.3	Candida albicans SC5314	likely cytosolic ribosomal protein L16
12	1	1	11.99	4	5.85E+04	22692.3	Candida albicans SC5314	likely cytosolic ribosomal protein L16
12	1	1	11.99	4	5.85E+04	22692.3	Candida albicans SC5314	likely cytosolic ribosomal protein L16
12	1	1	11.99	4	5.85E+04	22677.3	Candida parapsilosis	hypothetical protein CPAR2_109630
12	1	1	11.99	4	5.85E+04	22458	Lachancea thermotolerans CBS 6340	KLTH0G09042p
12	1	1	11.99	4	5.85E+04	22458	Lachancea thermotolerans	60S ribosomal protein L16
12	1	1	11.99	4	5.85E+04	22338.8	Kluyveromyces lactis	KLLA0F04675p
12	1	1	11.99	4	5.85E+04	22338.8	Kluyveromyces lactis NRRL Y-1140	60S ribosomal protein L16
12	1	1	11.99	4	5.85E+04	22269.8	Saccharomyces cerevisiae YJM789	ribosomal protein L16A
12	1	1	11.99	4	5.85E+04	22269.8	Saccharomyces cerevisiae CEN.PK113-7D	Rpl16ap
12	1	1	11.99	4	5.85E+04	22248.7	Saccharomyces cerevisiae EC1118	Rpl16bp
12	1	1	11.99	4	5.85E+04	22248.7	Saccharomyces cerevisiae S288c	TPA: ribosomal 60S subunit protein L16B
12	1	1	11.99	4	5.85E+04	22248.7	Saccharomyces cerevisiae S288c	ribosomal 60S subunit protein L16B
12	1	1	11.99	4	5.85E+04	22248.7	Saccharomyces cerevisiae	unknown
12	1	1	11.99	4	5.85E+04	22241.8	Saccharomyces cerevisiae EC1118	Rpl16ap
12	1	1	11.99	4	5.85E+04	22218.7	Saccharomyces cerevisiae YJM789	ribosomal protein L16B
12	1	1	11.99	4	5.85E+04	22218.7	Saccharomyces cerevisiae Kyokai no. 7	K7_Rpl16bp
12	1	1	11.99	4	5.85E+04	22204.7	Saccharomyces cerevisiae CEN.PK113-7D	Rpl16bp
12	1	1	11.99	4	5.85E+04	22200.7	Saccharomyces cerevisiae S288c	TPA: ribosomal 60S subunit protein L16A
12	1	1	11.99	4	5.85E+04	22200.7	Saccharomyces cerevisiae Kyokai no. 7	K7_Rpl16ap
12	1	1	11.99	4	5.85E+04	22200.7	Saccharomyces cerevisiae S288c	ribosomal 60S subunit protein L16A RecName: Full=60S ribosomal protein L16-A; AltName: Full=L13a; AltName: Full=L21; AltName: Full=RP22; AltName: Full=YL15
12	1	1	11.99	4	5.85E+04	22200.7	YEAST	ribosomal 60S subunit protein L16A RecName: Full=60S ribosomal protein L16-A; AltName: Full=L13a; AltName: Full=L21; AltName: Full=RP22; AltName: Full=YL15
12	1	1	11.99	7	5.85E+04	12631.6	Saccharomyces cerevisiae	unknown

13	1	1	11.9	4.1	6.03E+04	21713.8	Saccharomyces cerevisiae	putative second copy of ribosomal protein gene YL9A, SWISS_PROT:RL9_YEAST
Totals:	92	85						

Appendix 9: Transposable element predictions

Table 1: DFAM Transposable element predictions for the genome of *E. bieneusi*

Type	Superfamily	# target	name	query name	ali-st	ali-en	description of target
Retrotransposon	ERV1	ERV1-N6-I_DR	ABGB01000079	1716	1195	Internal sequence of a non-autonomous endogenous retrovirus from zebrafish	
Retrotransposon	ERV1	ERV1-N6-I_DR	ABGB01000088	1503	971	Internal sequence of a non-autonomous endogenous retrovirus from zebrafish	
Retrotransposon	ERV1	ERV1-N6-I_DR	ABGB01000090	1085	1653	Internal sequence of a non-autonomous endogenous retrovirus from zebrafish	
Retrotransposon	ERV1	ERV1-N6-I_DR	ABGB01000148	584	1031	Internal sequence of a non-autonomous endogenous retrovirus from zebrafish	
Retrotransposon	ERV1	ERV1-N7-I_DR	ABGB01000086	1655	2251	Internal sequence of a non-autonomous endogenous retrovirus from zebrafish	
Retrotransposon	ERV1	ERV1-N7-I_DR	ABGB01000148	993	1449	Internal sequence of a non-autonomous endogenous retrovirus from zebrafish	
Retrotransposon	ERV1	ZFERV-2N1-I_DR	ABGB01000280	107	724	LTR retrotransposon from zebrafish: internal sequence	
Retrotransposon	ERVK	ERVB5_1-I_MM	ABGB01000252	1689	1292	Mouse endogenous beta retrovirus ERVB5_1, internal sequence.	
Retrotransposon	ERVK	ETnERV3	ABGB01000211	68	391	ERV2 Endogenous Retrovirus from mouse.	
Retrotransposon	ERVK	ETnERV3	ABGB01000367	999	1227	ERV2 Endogenous Retrovirus from mouse.	
Retrotransposon	gypsy	Gypsy-221_DR-LTR	ABGB01000071	1929	2281	LTR retrotransposon from zebrafish: long terminal repeat	
Retrotransposon	gypsy	Gypsy-221_DR-LTR	ABGB01000083	373	701	LTR retrotransposon from zebrafish: long terminal repeat	
Retrotransposon	gypsy	Gypsy-221_DR-LTR	ABGB01000083	1327	915	LTR retrotransposon from zebrafish: long terminal repeat	
Retrotransposon	gypsy	Gypsy-221_DR-LTR	ABGB01000146	491	61	LTR retrotransposon from zebrafish: long terminal repeat	
Retrotransposon	gypsy	Gypsy-221_DR-LTR	ABGB01000177	384	195	LTR retrotransposon from zebrafish: long terminal repeat	
Retrotransposon	gypsy	Gypsy-221_DR-LTR	ABGB01000182	526	297	LTR retrotransposon from zebrafish: long terminal repeat	
Retrotransposon	gypsy	Gypsy-221_DR-LTR	ABGB01000183	275	19	LTR retrotransposon from zebrafish: long terminal repeat	
Retrotransposon	gypsy	Gypsy-221_DR-LTR	ABGB01000262	814	1127	LTR retrotransposon from zebrafish: long terminal repeat	
Retrotransposon	gypsy	Gypsy-221_DR-LTR	ABGB01000262	1229	931	LTR retrotransposon from zebrafish: long terminal repeat	
Retrotransposon	gypsy	Gypsy-221_DR-LTR	ABGB01000489	1244	906	LTR retrotransposon from zebrafish: long terminal repeat	
Retrotransposon	gypsy	Gypsy-221_DR-LTR	ABGB01000543	357	576	LTR retrotransposon from zebrafish: long terminal repeat	
Retrotransposon	gypsy	Gypsy77-I_DR	ABGB01000167	175	431	LTR retrotransposon from zebrafish: internal sequence	
Retrotransposon	gypsy	Gypsy77-I_DR	ABGB01000192	1778	1377	LTR retrotransposon from zebrafish: internal sequence	
Retrotransposon	gypsy	Gypsy91-I_DR	ABGB01000065	2081	1858	LTR retrotransposon from zebrafish: internal sequence	
Retrotransposon	I	Nimb-16_DR	ABGB01000079	1471	1029	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000083	1545	2127	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000111	809	102	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000144	130	603	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000146	361	727	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000167	33	490	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000168	553	872	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000198	1458	1071	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000218	491	47	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000253	36	407	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000262	1670	1059	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000270	1577	1159	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000278	1645	1214	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000283	1557	1110	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	Nimb-16_DR	ABGB01000563	378	919	Non-LTR retrotransposon from zebrafish	
Retrotransposon	I	NonLTR-1 DR	ABGB01000231	1649	1273	Nonautonomous retrotransposable element from zebrafish	
Retrotransposon	L1	L1-33_DR	ABGB01000111	4	169	Non-LTR retrotransposon from zebrafish	
Retrotransposon	L1	L1-33 DR	ABGB01000177	119	266	Non-LTR retrotransposon from zebrafish	
Retrotransposon	L1	L1-84_DR	ABGB01000144	861	265	Non-LTR retrotransposon from zebrafish	
Retrotransposon	L1	L1-84_DR	ABGB01000231	1121	1594	Non-LTR retrotransposon from zebrafish	

Retrotransposon Ngaro	DIRS-N2_DR	ABGB01000147	219	634	DIRS-type LTR retrotransposon from zebrafish
Retrotransposon Ngaro	DIRS-N2_DR	ABGB01000167	713	278	DIRS-type LTR retrotransposon from zebrafish
Retrotransposon Ngaro	DIRS-N2_DR	ABGB01000246	1578	1254	DIRS-type LTR retrotransposon from zebrafish
Retrotransposon Ngaro	DIRS-N2_DR	ABGB01000513	605	319	DIRS-type LTR retrotransposon from zebrafish
Unknown	Undefined	TE-X-4_DR	ABGB01000071	2273	1888Non-autonomous transposable element from zebrafish
Unknown	Undefined	TE-X-4_DR	ABGB01000082	1207	1785Non-autonomous transposable element from zebrafish
Unknown	Undefined	TE-X-4_DR	ABGB01000083	427	1002Non-autonomous transposable element from zebrafish
Unknown	Undefined	TE-X-4_DR	ABGB01000177	305	603Non-autonomous transposable element from zebrafish
Unknown	Undefined	TE-X-4_DR	ABGB01000183	337	724Non-autonomous transposable element from zebrafish
Unknown	Undefined	TE-X-4_DR	ABGB01000196	1385	1754Non-autonomous transposable element from zebrafish
Unknown	Undefined	TE-X-4_DR	ABGB01000236	1202	1677Non-autonomous transposable element from zebrafish
Unknown	Undefined	TE-X-4_DR	ABGB01000282	574	6Non-autonomous transposable element from zebrafish
Unknown	Undefined	TE-X-4_DR	ABGB01000289	236	756Non-autonomous transposable element from zebrafish
Unknown	Undefined	TE-X-4_DR	ABGB01001018	483	1104Non-autonomous transposable element from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000091	1835	1502Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000115	1141	1744Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000135	666	485Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000135	949	582Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000146	795	383Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000150	1238	1662Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000166	679	466Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000168	697	312Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000182	435	182Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000192	1364	1493Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000198	1238	1548Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000210	108	509Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000212	370	604Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000228	395	730Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000228	396	3Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000228	629	384Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000232	38	613Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000236	1699	1468Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000242	1019	1452Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000244	1404	1654Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000252	1636	1156Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000253	126	604Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000261	1347	1536Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000292	1130	1353Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000292	1616	1175Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000295	65	501Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000316	210	611Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000320	319	37Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000330	3	433Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000378	1151	1468Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000394	1240	1460Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000460	927	1306Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000460	1361	955Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000535	220	477Non-LTR retrotransposon from zebrafish

Unknown	Undefined	Tx1-20_DR	ABGB01000535	799	410Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000543	518	57Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01000636	1247	884Non-LTR retrotransposon from zebrafish
Unknown	Undefined	Tx1-20_DR	ABGB01001143	581	958Non-LTR retrotransposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000059	18	456DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000071	1133	2120DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000081	1370	452DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000083	1082	785DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000086	2097	1754DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000088	1679	1265DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000090	1693	2029DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000091	1175	640DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000091	1701	1211DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000093	1492	580DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000135	275	637DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000147	332	829DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000148	1119	468DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000184	1581	1887DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000202	1248	1686DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000212	191	483DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000222	1133	1579DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000226	1760	1315DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000234	504	20DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000246	1338	1690DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000252	1328	1684DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000273	70	503DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000280	51	444DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000292	1153	1620DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000335	1495	1167DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000368	315	811DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000368	573	124DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000445	571	6DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-10N1_DR	ABGB01000686	696	1176DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-N10_DR	ABGB01000079	1878	1453DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-N17_DR	ABGB01000242	1539	1197DNA transposon from zebrafish
Transposon	Crypton-H	LRS_DR	ABGB01000065	1935	1776Crypton DNA transposon from zebrafish
Transposon	Crypton-H	LRS_DR	ABGB01000065	2319	2007Crypton DNA transposon from zebrafish
Transposon	Crypton-H	LRS_DR	ABGB01000083	1455	1786Crypton DNA transposon from zebrafish
Transposon	Crypton-H	LRS_DR	ABGB01000091	911	1370Crypton DNA transposon from zebrafish
Transposon	Crypton-H	LRS_DR	ABGB01000152	773	115Crypton DNA transposon from zebrafish
Transposon	Crypton-H	LRS_DR	ABGB01000236	1318	1734Crypton DNA transposon from zebrafish
Transposon	Crypton-H	LRS_DR	ABGB01000513	341	628Crypton DNA transposon from zebrafish
Transposon	Crypton-H	LRS_DR	ABGB01000513	557	26Crypton DNA transposon from zebrafish
Transposon	Dada	Dada-tA_DR	ABGB01000088	1905	1378Dada-type DNA transposon from zebrafish
Transposon	Dada	Dada-U6_DR	ABGB01000065	848	1581Dada-type DNA transposon from zebrafish
Transposon	Dada	Dada-U6_DR	ABGB01000065	1474	1098Dada-type DNA transposon from zebrafish
Transposon	Dada	Dada-U6_DR	ABGB01000083	1509	1037Dada-type DNA transposon from zebrafish
Transposon	Dada	Dada-U6_DR	ABGB01000088	1284	839Dada-type DNA transposon from zebrafish

Transposon	Dada	Dada-U6_DR	ABGB01000091	1623	1055Dada-type DNA transposon from zebrafish
Transposon	Dada	Dada-U6_DR	ABGB01000175	1197	1794Dada-type DNA transposon from zebrafish
Transposon	Dada	Dada-U6_DR	ABGB01000286	1614	1008Dada-type DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000082	1625	1314DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000147	93	455DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000166	487	123DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000166	741	501DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000182	604	257DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000183	294	626DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000185	761	210DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000200	389	10DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000202	1472	1718DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000202	1663	1322DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000204	1818	1518DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000211	178	682DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000212	399	701DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000218	594	201DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000220	1731	1392DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000231	1652	1144DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000240	482	10DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000262	1185	839DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000367	1407	1125DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000368	280	791DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01000432	1399	1035DNA transposon from zebrafish
Transposon	Ginger	Ginger1-1_DR	ABGB01001146	654	944DNA transposon from zebrafish
Transposon	hAT	hAT-N36C_DR	ABGB01000228	267	556Putative non-autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N53_DR	ABGB01000059	147	583Putative non-autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N53_DR	ABGB01000083	1306	1625Putative non-autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N53_DR	ABGB01000166	592	246Putative non-autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N53_DR	ABGB01000177	195	34Putative non-autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N53_DR	ABGB01001146	754	1040Putative non-autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000027	344	752Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000081	975	1628Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000086	2411	1971Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000093	1340	1873Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000135	7	336Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000135	335	657Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000175	1764	1110Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000185	50	588Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000215	1764	1034Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000221	3	576Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000236	1528	1082Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000255	30	577Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000289	630	93Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000364	20	715Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000371	1488	1074Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000489	1238	759Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01000535	319	603Autonomous hAT DNA transposon from zebrafish

Transposon	hAT	hAT-N82_DR	ABGB01000733	882	611	Autonomous hAT DNA transposon from zebrafish
Transposon	hAT	hAT-N82_DR	ABGB01001146	986	707	Autonomous hAT DNA transposon from zebrafish
Transposon	Helitron	Helitron-N3_DR	ABGB01000906	1167	811	Non-autonomous Helitron from zebrafish
Transposon	Helitron	Helitron-N3_DR	ABGB01001018	419	975	Non-autonomous Helitron from zebrafish
Transposon	Maverick	Polinton-2N1_DR	ABGB01000137	1465	1079	Non-autonomous Maverick/Polinton DNA transposon from zebrafish
Transposon	Maverick	Polinton-2N1_DR	ABGB01000162	1893	1088	Non-autonomous Maverick/Polinton DNA transposon from zebrafish
Transposon	Maverick	Polinton-2N1_DR	ABGB01000183	413	103	Non-autonomous Maverick/Polinton DNA transposon from zebrafish
Transposon	Maverick	Polinton-2N1_DR	ABGB01000198	1094	1541	Non-autonomous Maverick/Polinton DNA transposon from zebrafish
Transposon	Maverick	Polinton-2N1_DR	ABGB01000244	1054	1542	Non-autonomous Maverick/Polinton DNA transposon from zebrafish
Transposon	Maverick	Polinton-2N1_DR	ABGB01000316	136	613	Non-autonomous Maverick/Polinton DNA transposon from zebrafish
Transposon	Maverick	Polinton-2N1_DR	ABGB01000441	164	609	Non-autonomous Maverick/Polinton DNA transposon from zebrafish
Transposon	PIF Harbinger	Harbinger-2_DR	ABGB01000135	840	128	DNA transposon from zebrafish
Transposon	PiggyBac	piggyBac-N7_DR	ABGB01000183	417	84	DNA transposon from zebrafish
Transposon	TcMar	Mariner-N5_DR	ABGB01000071	2215	1840	Tc1/Mariner-type non-autonomous DNA transposon from zebrafish
Transposon	TcMar	Mariner-N5_DR	ABGB01000222	1245	1712	Tc1/Mariner-type non-autonomous DNA transposon from zebrafish
Transposon	TcMar	Mariner-N5_DR	ABGB01001143	1148	890	Tc1/Mariner-type non-autonomous DNA transposon from zebrafish
Transposon	TcMar	Mariner-N5B_DR	ABGB01000027	88	503	Tc1/Mariner-type non-autonomous DNA transposon from zebrafish
Transposon	TcMar	Mariner-N5B_DR	ABGB01000065	1196	1797	Tc1/Mariner-type non-autonomous DNA transposon from zebrafish
Transposon	TcMar	Mariner-N5B_DR	ABGB01000086	2190	1571	Tc1/Mariner-type non-autonomous DNA transposon from zebrafish
Transposon	TcMar	Mariner-N5B_DR	ABGB01000111	142	870	Tc1/Mariner-type non-autonomous DNA transposon from zebrafish
Transposon	TcMar	Mariner-N5B_DR	ABGB01000168	3	765	Tc1/Mariner-type non-autonomous DNA transposon from zebrafish
Transposon	TcMar	Mariner-N5B_DR	ABGB01000204	1691	1172	Tc1/Mariner-type non-autonomous DNA transposon from zebrafish
Transposon	TcMar	Mariner-N5B_DR	ABGB01000242	1621	1317	Tc1/Mariner-type non-autonomous DNA transposon from zebrafish
Transposon	TcMar	Mariner-N5B_DR	ABGB01000563	723	29	Tc1/Mariner-type non-autonomous DNA transposon from zebrafish
Transposon	TcMar-Tc4	TC4	ABGB01000329	435	124	Transposable element Tc4.
Transposon	Undefined	DNA2-10_DR	ABGB01000273	40	459	DNA transposon from zebrafish
Transposon	Undefined	DNA7-N1_DR	ABGB01000093	1512	707	DNA transposon from zebrafish
Transposon	Unknown	Chaplin1A_DR	ABGB01000167	588	159	Non-autonomous DNA transposon from zebrafish

Table 2: DFAM transposable element predictions for the genome of *Ent. canceri*

Type	superfamily	target name	query name	ali-st	ali-en	description of target
Retrotransposon	CR1	CR1_Mam	scaffold_1592	626	111	Mammalian CR1 (Chicken Repeat 1) LINE
Retrotransposon	CR1	CR1_Mam	scaffold_172	1907	660	Mammalian CR1 (Chicken Repeat 1) LINE
Retrotransposon	CR1	CR1_Mam	scaffold_585	223	1028	Mammalian CR1 (Chicken Repeat 1) LINE
Retrotransposon	CR1	CR1-1_Amn	scaffold_888	2	303	CR1 (Chicken Repeat 1) retrotransposon, CR1-1_Amn subfamily
Retrotransposon	CR1	CR1-3_Croc	scaffold_604	28	595	CR1 (Chicken Repeat 1) retrotransposon, CR1-3_Croc subfamily
Retrotransposon	CR1	CR1-L3A_Croc	scaffold_1191	767	572	CR1 (Chicken Repeat 1) retrotransposon, CR1-L3A_Croc subfamily
Retrotransposon	CR1	CR1-L3A_Croc	scaffold_121	2729	4256	CR1 (Chicken Repeat 1) retrotransposon, CR1-L3A_Croc subfamily
Retrotransposon	CR1	CR1-L3A_Croc	scaffold_1291	80	715	CR1 (Chicken Repeat 1) retrotransposon, CR1-L3A_Croc subfamily
Retrotransposon	CR1	CR1-L3A_Croc	scaffold_1475	662	139	CR1 (Chicken Repeat 1) retrotransposon, CR1-L3A_Croc subfamily
Retrotransposon	CR1	CR1-L3A_Croc	scaffold_149	1561	22	CR1 (Chicken Repeat 1) retrotransposon, CR1-L3A_Croc subfamily
Retrotransposon	CR1	CR1-L3A_Croc	scaffold_1692	2	492	CR1 (Chicken Repeat 1) retrotransposon, CR1-L3A_Croc subfamily
Retrotransposon	CR1	CR1-L3A_Croc	scaffold_493	8	415	CR1 (Chicken Repeat 1) retrotransposon, CR1-L3A_Croc subfamily
Retrotransposon	CR1	CR1-L3A_Croc	scaffold_976	822	4	CR1 (Chicken Repeat 1) retrotransposon, CR1-L3A_Croc subfamily
Retrotransposon	CR1	L3b_3end	scaffold_276	7	325	CR1 (Chicken Repeat 1) retrotransposon, L3b_3end subfamily
Retrotransposon	CR1	Plat_L3	scaffold_490	308	1270	CR1 (Chicken Repeat 1) retrotransposon, Plat_L3 subfamily
Retrotransposon	CR1	X2_LINE	scaffold_149	1064	2	CR1 (Chicken Repeat 1) retrotransposon, X2_LINE subfamily
Retrotransposon	CR1	X2_LINE	scaffold_294	549	125	CR1 (Chicken Repeat 1) retrotransposon, X2_LINE subfamily

Retrotransposon	CR1	X2_LINE	scaffold_997	397	20	CR1 (Chicken Repeat 1) retrotransposon, X2_LINE subfamily
Retrotransposon	CR1	X21_LINE	scaffold_1191	707	498	Repetitive element conserved in all mammals.
Retrotransposon	CR1	X21_LINE	scaffold_888	167	376	Repetitive element conserved in all mammals.
Retrotransposon	CR1	X6A_LINE	scaffold_123	4191	3609	CR1 (Chicken Repeat 1) retrotransposon, X6A_LINE subfamily
Retrotransposon	CR1	X6A_LINE	scaffold_698	13	595	CR1 (Chicken Repeat 1) retrotransposon, X6A_LINE subfamily
Retrotransposon	CR1	X6A_LINE	scaffold_922	885	35	CR1 (Chicken Repeat 1) retrotransposon, X6A_LINE subfamily
Retrotransposon	CR1	Plat_L3	scaffold_714	6	854	CR1 (Chicken Repeat 1) retrotransposon, Plat_L3 subfamily
Retrotransposon	CR1	Plat_L3	scaffold_997	826	140	CR1 (Chicken Repeat 1) retrotransposon, Plat_L3 subfamily
Retrotransposon	ERVK	RLTR20A4	scaffold_532	1197	1063	Long terminal repeat of retrovirus-like element.
Retrotransposon	gypsy	ACCORD_I	scaffold_21	15248	14202	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	scaffold_42	18665	19755	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	scaffold_60	3395	4648	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	scaffold_99	2459	3644	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	Gypsy-117-I_DR	scaffold_343	999	694	LTR retrotransposon from zebrafish: internal sequence
Retrotransposon	gypsy	Gypsy-117-I_DR	scaffold_425	734	1040	LTR retrotransposon from zebrafish: internal sequence
Retrotransposon	gypsy	Gypsy-117-I_DR	scaffold_430	1041	1346	LTR retrotransposon from zebrafish: internal sequence
Retrotransposon	gypsy	Gypsy-117-I_DR	scaffold_559	871	552	LTR retrotransposon from zebrafish: internal sequence
Retrotransposon	gypsy	Gypsy-216_DR-I	scaffold_1367	686	24	LTR retrotransposon from zebrafish: internal sequence
Retrotransposon	gypsy	Gypsy-216_DR-I	scaffold_204	437	1511	LTR retrotransposon from zebrafish: internal sequence
Retrotransposon	gypsy	Gypsy-216_DR-I	scaffold_554	1196	915	LTR retrotransposon from zebrafish: internal sequence
Retrotransposon	gypsy	Gypsy-216_DR-I	scaffold_991	831	315	LTR retrotransposon from zebrafish: internal sequence
Retrotransposon	gypsy	Gypsy-231_DR-LTR	scaffold_444	887	757	LTR retrotransposon from zebrafish: long terminal repeat
Retrotransposon	gypsy	Gypsy121-I_DR	scaffold_1319	147	17	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy121-I_DR	scaffold_559	641	8	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy121-I_DR	scaffold_819	329	944	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy127-I_DR	scaffold_268	1336	1174	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy127-I_DR	scaffold_365	636	798	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy127-I_DR	scaffold_397	620	782	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy127-I_DR	scaffold_485	75	237	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy127-I_DR	scaffold_596	440	602	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy127-I_DR	scaffold_837	208	46	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy127-I_DR	scaffold_929	581	399	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy44-I_DR	scaffold_219	1067	1280	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy44-I_DR	scaffold_357	164	47	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy44-I_DR	scaffold_368	1279	1492	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy44-I_DR	scaffold_407	822	1126	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy44-I_DR	scaffold_436	282	69	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy44-I_DR	scaffold_555	688	373	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy44-I_DR	scaffold_765	230	539	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy44-I_DR	scaffold_840	258	45	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy44-I_DR	scaffold_893	896	606	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_1003	646	6	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_1064	239	628	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_1319	709	388	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_163	28	185	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_163	2132	2522	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_204	1749	2304	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_219	1068	1379	Internal sequence of a Gypsy LTR retrotransposon from zebrafish

Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_351	1243	316	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_403	14	522	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_446	5	413	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_448	1274	2	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_508	589	200	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_554	625	70	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy45-I_DR	scaffold_747	35	708	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Gypsy47-I_DR	scaffold_819	178	764	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	QUASIMODO_I	scaffold_171	1151	1973	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO2-I_DM	scaffold_42	18526	19059	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	ZAM_I	scaffold_21	15266	14983	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	scaffold_42	18512	19057	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	I	Nimb-13_DR	scaffold_364	667	1357	Non-LTR retrotransposon from zebrafish
Retrotransposon	ID	ID2	scaffold_21	32563	32506	ID2 subfamily
Retrotransposon	ID	ID2	scaffold_21	32863	32806	ID2 subfamily
Retrotransposon	ID	ID2	scaffold_271	41	98	ID2 subfamily
Retrotransposon	JOCKEY	FW2_DM	scaffold_91	3627	4163	FW2_DM is a non-LTR retrotransposon
Transposon	CMC-EnSpm	EnSpm-7_DR	scaffold_151	333	571	DNA transposon from zebrafish
Transposon	CMC-EnSpm	EnSpm-N13_DR	scaffold_400	627	407	DNA transposon from zebrafish
Transposon	hAT-Ac	hAT-N131_DR	scaffold_151	661	499	DNA transposon from zebrafish
Transposon	Merlin	DNA8-21_DR	scaffold_416	1235	985	DNA transposon from zebrafish
Transposon	Merlin	Merlin1_HS	scaffold_125	2313	1903	Merlin1_HS -- Merlin DNA transposon
Transposon	Merlin	Merlin1_HS	scaffold_156	1026	1426	Merlin1_HS -- Merlin DNA transposon
Transposon	Merlin	Merlin1_HS	scaffold_207	369	766	Merlin1_HS -- Merlin DNA transposon
Transposon	Merlin	Merlin1_HS	scaffold_275	1477	1083	Merlin1_HS -- Merlin DNA transposon
Transposon	Merlin	Merlin1_HS	scaffold_80	7495	7892	Merlin1_HS -- Merlin DNA transposon
Transposon	Merlin	Merlin1_HS	scaffold_953	21	431	Merlin1_HS -- Merlin DNA transposon
Transposon	PIF-Harbinger	Harbinger-2_DR	scaffold_1105	258	450	DNA transposon from zebrafish
Transposon	PIF-Harbinger	Harbinger-2_DR	scaffold_77	245	353	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	scaffold_167	1850	1567	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	scaffold_60	8501	8712	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8_DR	scaffold_227	2481	2833	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8_DR	scaffold_597	507	51	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	scaffold_101	3993	4323	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	scaffold_102	584	180	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	scaffold_102	940	615	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	scaffold_124	1588	1266	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	scaffold_159	14	264	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	scaffold_227	1934	2259	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	scaffold_227	2290	2693	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	scaffold_597	674	196	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	scaffold_60	8286	8713	DNA transposon from zebrafish

Table 3: DFAM transposable element predictions for the genome of *E. hepatopenaei*

Class	superfamily	target name	query name	all-st	all-en	description of target
Retrotransposon	gypsy	ACCORD_I	lc EHP_503	2419	2107	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	Gypsy4_I	lc EHP_549	2022	2253	GYPSY4_I is an internal portion of the GYPSY4 endogenous retrovirus
Retrotransposon	gypsy	QUASIMODO2-I_DM	lc EHP_549	2453	3049	Drosophila melanogaster retroviral like element QUASIMODO2, internal.

Retrotransposon	gypsy	ZAM_I	lc EHP_503	2419	2010	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	ID	BC1_Mm	lc EHP_223	1786	1726	Mouse neural specific BC1 RNA and ID repetitive sequence.
Retrotransposon	L1	L1M3_orf2	lc EHP_830	1848	2084	ORF2 from L1 retrotransposon, L1M3_orf2 subfamily
Transposon	Merlin	Merlin1_HS	lc EHP_312	4850	4443	Merlin1_HS -- Merlin DNA transposon
Transposon	Sola	TDR17B	lc EHP_392	420	272	Non-autonomous DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-1_DR	lc EHP_106	8575	8734	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lc EHP_135	2383	1976	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lc EHP_61	18926	0	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lc EHP_137	3980	3469	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lc EHP_1444	830	1489	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lc EHP_37	10722	0	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lc EHP_61	19001	8	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lc EHP_70	17510	8	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-N6_DR	lc EHP_137	4008	3698	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	S_DM	lc EHP_135	8729	8288	S_DM is a Tc1/mariner-like DNA transposon.
Transposon	TcMar-Tc1	Tc1-8B_DR	lc EHP_135	8876	8465	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lc EHP_189	4099	3752	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lc EHP_37	10848	1	DNA transposon from zebrafish
Transposon	Undefined	MSAT-3_DR	lc EHP_83	22187	0	Minisatellite-like DNA
Transposon	Unknown	Tc3	lc EHP_106	12981	9	An active DNA transposon - a consensus.
Transposon	Unknown	Tc3	lc EHP_106	18189	1	An active DNA transposon - a consensus.
Transposon	Unknown	Tc3	lc EHP_135	9711	9856	An active DNA transposon - a consensus.
Transposon	Unknown	Tc3	lc EHP_4818	234	89	An active DNA transposon - a consensus.
Transposon	Unknown	Tc3	lc EHP_64	5998	6413	An active DNA transposon - a consensus.

Table 4: DFAM transposable element predictions for the genome of *H. eriocheir canceri*

Type	superfamily	target name	query name	all-st	ali-en	description of target
Retrotransposon	gypsy	ACCORD_I	lc scaffold_102_5	117	474	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	lc scaffold_107_1	1322	248	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	lc scaffold_315	3164	3839	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	lc scaffold_389_0	10	363	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	lc scaffold_432	1735	672	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	lc scaffold_459	1126	1517	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	lc scaffold_604	1699	515	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	lc scaffold_628	2009	1636	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	lc scaffold_80	4212	5347	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	lc scaffold_862	66	421	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	BURDOCK_I	lc scaffold_299	1364	1679	Drosophila melanogaster BURDOCK retrotransposon - internal sequence.
Retrotransposon	gypsy	BURDOCK_I	lc scaffold_459	1269	2321	Drosophila melanogaster BURDOCK retrotransposon - internal sequence.
Retrotransposon	gypsy	GTWIN_I	lc scaffold_862	25	145	GTWIN_I is an internal part of Gypsy-related retrovirus GTWIN.
Retrotransposon	gypsy	Gypsy4_I	lc scaffold_127_2	615	143	GYPY4_I is an internal portion of the GYPY4 endogenous retrovirus
Retrotransposon	gypsy	Gypsy4_I	lc scaffold_178	43	287	GYPY4_I is an internal portion of the GYPY4 endogenous retrovirus
Retrotransposon	gypsy	Gypsy4_I	lc scaffold_346	2129	2597	GYPY4_I is an internal portion of the GYPY4 endogenous retrovirus
Retrotransposon	gypsy	Gypsy4_I	lc scaffold_782	700	229	GYPY4_I is an internal portion of the GYPY4 endogenous retrovirus
Retrotransposon	gypsy	Gypsy4_I	lc scaffold_844	1734	1263	GYPY4_I is an internal portion of the GYPY4 endogenous retrovirus
Retrotransposon	gypsy	Gypsy6-I_DR	lc scaffold_267	2647	3013	Internal sequence of a Gypsy LTR retrotransposon from zebrafish
Retrotransposon	gypsy	Invader6_I	lc scaffold_178_6	31	529	INVADER6_I is an internal portion of the INVADER6 endogenous retrovirus.
Retrotransposon	gypsy	Invader6_I	lc scaffold_285	2480	1929	INVADER6_I is an internal portion of the INVADER6 endogenous retrovirus.

Retrotransposon	gypsy	QUASIMODO_I	lcl scaffold_13	7944	6962	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	lcl scaffold_13	12933	1189	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	lcl scaffold_168	4044	5115	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	lcl scaffold_212	2565	3446	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	lcl scaffold_262	648	1696	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	lcl scaffold_267	2739	3807	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	lcl scaffold_280	1843	2522	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	lcl scaffold_285	2292	1340	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	lcl scaffold_33	6951	7359	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	lcl scaffold_347	2109	1156	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	lcl scaffold_38	1285	2376	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	lcl scaffold_879	763	435	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO2-I_DM	lcl scaffold_241	2092	925	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	QUASIMODO2-I_DM	lcl scaffold_276	1751	2776	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	QUASIMODO2-I_DM	lcl scaffold_28	8978	1013	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	QUASIMODO2-I_DM	lcl scaffold_28	4994	4037	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	QUASIMODO2-I_DM	lcl scaffold_29	9980	9326	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	QUASIMODO2-I_DM	lcl scaffold_305	2356	1205	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	QUASIMODO2-I_DM	lcl scaffold_346	2032	2443	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	QUASIMODO2-I_DM	lcl scaffold_407	4	292	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	QUASIMODO2-I_DM	lcl scaffold_604	1701	503	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	QUASIMODO2-I_DM	lcl scaffold_80	1718	714	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	TIRANT_I	lcl scaffold_743	1134	903	An internal portion of Gypsy-like LTR retrotransposon Tirant
Retrotransposon	gypsy	ZAM_I	lcl scaffold_127	5757	6435	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_133	1481	2368	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_178	6	50	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_180	1692	2748	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_188	4724	4297	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_199	1	283	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_280	1775	2114	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_315	3119	3842	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_334	2	573	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_346	2114	2571	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_38	1311	2330	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_432	1879	685	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_570	1042	40	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_64	5715	6824	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_814	120	932	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_870	869	552	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	lcl scaffold_874	859	6	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	Unknown	MDG3_I	lcl scaffold_347	2295	1856	Internal part of D.melanogaster MDG3 retrotransposon.
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_102	7	945	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_103	2	796	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_110	7	731	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_116		302	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_117		2987	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_123	9	732	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_127		1803	Tc1/Mariner-type DNA transposon from zebrafish

Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_127_1	4	137	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_134_1	479	874	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_134_7	815	1126	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_154_6	3176	2866	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_162_8	567	256	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_166_7	438	758	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_169_9	6	162	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_177_4	357	669	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_195_4	627	1039	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_213_9	133	530	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_219_1	550	749	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_234_5	81	387	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_241_4	493	178	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_246_6	21	155	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_277_2	641	519	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_295_3	2439	2139	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_32_1	52	183	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_32_2	1719	1402	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_323_1	754	1059	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_328_1	3113	2798	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_345_1	2653	3034	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_354_1	2346	2040	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_365_1	2198	2515	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_462_1	1256	944	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_508_1	1992	1680	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_532_1	331	710	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_532_2	1998	1672	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_535_1	1507	1192	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_61_1	6972	6672	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_692_1	2	119	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_737_1	29	247	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	lcl scaffold_816_1	24	154	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_103_2	884	1192	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_108_9	929	369	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_111_1	402	27	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_112_4	82	356	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_112_8	1025	1325	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_117_6	6403	5851	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_125_6	15	387	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_137_6	136	667	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_144_6	1080	933	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_150_1	3699	3532	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_158_6	766	1288	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_162_6	573	267	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_175_1	2919	2617	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_175_2	4403	3886	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_184_1	2807	3106	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_188_1	107	545	Tc1/Mariner-type DNA transposon from zebrafish

Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_213 4	220	528	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_222 4	747	574	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_249	2541	1971	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_256	3467	3075	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_260 4	39	210	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_261 9	626	455	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_289	3633	3113	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_32	28	175	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_324 8	147	583	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_354	2357	2051	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_359	3438	3271	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_393	3134	2574	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_414 8	500	328	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_464	729	420	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_496	1300	739	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_515	2712	2474	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_553	401	924	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_567	579	867	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_6	14755	1516 8	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_648	908	1429	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_714	1698	2000	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_722	464	25	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_847	302	834	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_870	16	163	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	lcl scaffold_980	1269	684	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-1_DR	lcl scaffold_414 8	440	326	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-1_DR	lcl scaffold_870	51	165	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_102	4503	5018	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_102 6	1487	984	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_105 0	37	425	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_107 4	727	427	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_108 0	1173	1410	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_110 7	642	1035	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_112 8	1030	1342	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_114 1	1133	1351	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_114 8	280	37	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_116	5406	4884	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_118 3	516	830	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_125 6	4	334	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_133 8	547	1085	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_134 7	724	1116	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_135	488	17	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_139 1	394	967	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_142 3	436	921	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_145 7	714	1072	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_147 2	706	1064	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_150	5185	5591	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_165 3	650	775	DNA transposon from zebrafish

Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_180	4168	4659	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_182 1	871	572	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_182 3	450	141	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_189 7	14	262	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_195	641	1029	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_195	3573	3077	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_195 2	356	628	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_199	4337	4901	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_199 0	159	680	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_204 2	6	579	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_220 5	308	2	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_244 0	1	506	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_263 9	450	47	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_265	670	177	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_270	3893	3487	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_270 9	423	645	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_280 0	636	494	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_297 2	311	2	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_318 2	317	2	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_322	3093	3615	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_329	3229	3698	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_331 2	301	531	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_333 4	510	236	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_334	2063	1859	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_344 2	563	29	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_349	2934	2620	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_360 0	404	538	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_375	3304	3031	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_380	1980	2287	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_380 9	1	292	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_393	3136	2611	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_410	2574	2253	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_413	202	652	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_467	1703	2019	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_468	943	451	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_478	286	43	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_486	191	598	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_506	2202	2584	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_515	2713	2502	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_528	740	168	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_549	2320	2541	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_561	202	18	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_566	1947	2275	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_626	469	881	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_687	1878	1726	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_689	1266	694	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_707	1396	1889	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_751	474	864	DNA transposon from zebrafish

Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_765	1890	1407	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_78	925	1314	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_816	1566	1773	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_840	1078	1560	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_869	1107	1518	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_876	553	286	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_898	1026	624	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_978	1179	654	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	lcl scaffold_980	1272	704	DNA transposon from zebrafish
Transposon	TcMar-Tc1	TE-X-10_DR	lcl scaffold_6	14924	1525	Non-autonomous transposable element from zebrafish
Transposon	TcMar-Tc1	TZF28	lcl scaffold_1027	1040	725	Autonomous Tc1/Mariner-type DNA transposon from zebrafish
Transposon	Unknown	S2_DM	lcl scaffold_1050	205	515	S2_DM is a DNA transposon, a consensus sequence.
Transposon	Undefined	MSAT-3_DR	lcl scaffold_1819	461	619	Minisatellite-like DNA
Transposon	Undefined	MSAT-3_DR	lcl scaffold_891	296	461	Minisatellite-like DNA
Transposon	Undefined	MSAT-3_DR	lcl scaffold_891	407	591	Minisatellite-like DNA
Transposon	Undefined	MSAT-3_DR	lcl scaffold_891	536	695	Minisatellite-like DNA
Satellite	Undefined	MSAT-3_DR	lcl scaffold_1819	393	564	Minisatellite-like DNA

Table 5: DFAM transposable element predictions for the genome of *H. eriocheir*

Type	Superfamily	# target name	query name	all-st	all-en	description of target
Retrotransposon	gypsy	ACCORD_I	NODE_2927	830	1176	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	NODE_4215	4821	5721	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	ACCORD_I	NODE_4237	7524	6444	ACCORD_I is an internal portion of ACCORD retrovirus-like element
Retrotransposon	gypsy	BURDOCK_I	NODE_1615	2285	2538	Drosophila melanogaster BURDOCK retrotransposon - internal sequence.
Retrotransposon	gypsy	BURDOCK_I	NODE_2838	6002	7029	Drosophila melanogaster BURDOCK retrotransposon - internal sequence.
Retrotransposon	gypsy	BURDOCK_I	NODE_417	12577	13629	Drosophila melanogaster BURDOCK retrotransposon - internal sequence.
Retrotransposon	gypsy	BURDOCK_I	NODE_4895	4287	4580	Drosophila melanogaster BURDOCK retrotransposon - internal sequence.
Retrotransposon	gypsy	BURDOCK_I	NODE_8586	2000	2186	Drosophila melanogaster BURDOCK retrotransposon - internal sequence.
Retrotransposon	gypsy	Gypsy-19-I_DR	NODE_3128	9774	10127	Internal sequence of a Gypsy LTR retrotransposon from zebrafish - fossilize
Retrotransposon	gypsy	Gypsy4_I	NODE_1615	2586	2798	GYP4_I is an internal portion of the GYP4 endogenous retrovirus
Retrotransposon	gypsy	Gypsy4_I	NODE_6878	3923	3660	GYP4_I is an internal portion of the GYP4 endogenous retrovirus
Retrotransposon	gypsy	QUASIMODO_I	NODE_157	2318	1225	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	NODE_1655	285	104	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	NODE_2269	247	429	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	NODE_4895	21718	22397	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	NODE_5985	11828	11475	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO_I	NODE_892	426	744	QUASIMODO_I is an internal portion of QUASIMODO retrovirus-like element
Retrotransposon	gypsy	QUASIMODO2-I_DM	NODE_1319	20402	19612	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	QUASIMODO2-I_DM	NODE_3128	4313	3162	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	QUASIMODO2-I_DM	NODE_3833	5846	4647	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	QUASIMODO2-I_DM	NODE_3897	1475	574	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	QUASIMODO2-I_DM	NODE_5576	2099	2755	Drosophila melanogaster retroviral like element QUASIMODO2, internal.
Retrotransposon	gypsy	TIRANT_I	NODE_2927	845	1176	An internal portion of Gypsy-like LTR retrotransposon Tirant
Retrotransposon	gypsy	TIRANT_I	NODE_6878	3926	3615	An internal portion of Gypsy-like LTR retrotransposon Tirant
Retrotransposon	gypsy	ZAM_I	NODE_1192	712	161	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_1626	2942	3499	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_1793	30583	30902	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_2884	6618	7506	ZAM_I is an internal portion of the ZAM retrovirus.

Retrotransposon	gypsy	ZAM_I	NODE_2927	1744	2101	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_3080	2661	1551	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_3128	9868	10312	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_3558	12832	13150	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_4895	21651	22004	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_495	546	67	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_6878	4051	3724	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_6998	15259	15622	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_7395	4882	4419	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_776	7363	7732	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_777	33615	32566	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	gypsy	ZAM_I	NODE_892	533	904	ZAM_I is an internal portion of the ZAM retrovirus.
Retrotransposon	Unknown	MDG3_I	NODE_900	117	316	Internal part of D.melanogaster MDG3 retrotransposon.
Satellite	undefined	MSAT-3_DR	NODE_241	5896	6032	Minisatellite-like DNA
Satellite	undefined	MSAT-3_DR	NODE_241	5932	6103	Minisatellite-like DNA
Satellite	undefined	MSAT-3_DR	NODE_241	6129	6306	Minisatellite-like DNA
Satellite	undefined	MSAT-3_DR	NODE_241	6460	6602	Minisatellite-like DNA
Satellite	undefined	MSAT-3_DR	NODE_241	6517	6691	Minisatellite-like DNA
Satellite	undefined	MSAT-3_DR	NODE_241	6745	6930	Minisatellite-like DNA
Transposon	Merlin	Merlin1_HS	NODE_3286	655	1076	Merlin1_HS -- Merlin DNA transposon
Transposon	TcMar-Tc1	Tc1-5_DR	NODE_3559	344	222	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_1319	11305	10852	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_1353	956	1349	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_1438	27546	27053	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_1614	724	457	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_2449	63	195	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_2657	1208	1522	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_2715	1882	2419	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_2715	3495	3226	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_2823	3823	3330	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_2838	2074	2605	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_2838	2621	2771	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_2943	5281	5595	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_2949	4627	4103	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_3216	196	46	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_3224	17	294	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_3313	213	432	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_3444	1985	1685	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_369	340	503	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_415	237	88	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_4151	4945	4628	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_4215	8709	9046	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_4237	2074	2385	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_4299	2	306	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_4575	9	399	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_4812	2367	2669	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_4895	18576	18078	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_4925	4163	4484	DNA transposon from zebrafish

Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_492 5	34196	33879	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_502	9	398	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_533 8	2526	2743	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_549	262	13	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_573 6	215	10	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_579 6	422	823	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_598 3	6376	5983	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_601 6	1607	1923	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_616 8	624	326	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_636 6	350	32	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_650	4287	4770	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_666 9	1540	1921	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_666 9	4454	3949	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_744	22146	21759	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_770 5	942	1340	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_813 5	6759	7076	DNA transposon from zebrafish
Transposon	TcMar-Tc1	Tc1-8B_DR	NODE_965 6	486	1047	DNA transposon from zebrafish
Transposon	TE-X-10_DR	TE-X-10_DR	NODE_188 5	680	1014	Non-autonomous transposable element from zebrafish
Transposon	TE-X-10_DR	TE-X-10_DR	NODE_228 5	3603	3937	Non-autonomous transposable element from zebrafish
Transposon	Unknown	S2_DM	NODE_271 5	2188	2518	S2_DM is a DNA transposon, a consensus sequence.
Transposon	TcMar-Tc1	Mariner-10_DR	NODE_174	2	168	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_129 5	1044	728	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_135 3	1045	1351	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_146 6	606	209	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_160 8	5882	5562	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_161 1	24865	25178	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_161 3	18	238	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_161 6	44106	44421	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_162 6	7333	7639	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_228 5	917	1317	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_312 4	647	961	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_312 8	6625	7024	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_369 2	3941	3559	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_406	7	213	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_421 5	8788	9099	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_464 0	2549	2855	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_480 2	949	1248	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_482 0	618	210	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_492 5	691	383	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_554	661	341	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_636 5	10409	10712	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_674 1	4709	4306	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_751 1	8803	9193	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_796 9	1206	1596	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_872 9	6267	5956	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-12_DR	NODE_964 1	14379	14067	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_112 2	6827	7387	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_113 0	2219	1918	Tc1/Mariner-type DNA transposon from zebrafish

Transposon	TcMar-Tc1	Mariner-8_DR	NODE_119 2	3778	4084	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_129 7	459	150	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_152 9	139	567	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_160 1	1614	1133	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_188 5	401	925	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_228 5	3326	3852	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_283 8	3723	4143	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_299 2	18962	19467	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_312 8	6697	7016	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_355 8	6879	6358	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_355 9	381	234	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_370 4	2	254	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_383 3	12546	12847	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_423 7	2062	2369	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_423 7	13571	13175	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_443 4	4904	5169	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_526 1	3255	2847	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_548	289	48	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_560 1	5158	4884	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_598 3	6289	5982	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_606 1	70	598	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_636 5	10398	10701	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_636 5	29167	28795	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_644 9	5882	6185	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_666 9	1620	1922	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_744	22059	21758	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_777	37596	37905	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-8_DR	NODE_777	48462	48993	Tc1/Mariner-type DNA transposon from zebrafish
Transposon	TcMar-Tc1	Mariner-N20_DR	NODE_666 9	1667	1933	Tc1/Mariner-type DNA transposon from zebrafish

Table 6: REPEATFINDER transposable element predictions for the genome of *E. bieneusi*

Type	class/family	repeat	query	begin	end
Retrotransposons	LINE/CR1	X5B_LINE	ABGB01000014	115095	115127
Retrotransposons	LINE/CR1	L3	ABGB01000016	36535	36572
Retrotransposons	LINE/CR1	L3	ABGB01000023	25377	25454
Retrotransposons	LINE/CR1	X6B_LINE	ABGB01000176	29	120
Retrotransposons	LINE/CR1	L3	ABGB01000213	451	520
Retrotransposons	LINE/CR1	L3	ABGB01000234	1164	1233
Retrotransposons	LINE/CR1	L3	ABGB01000922	580	649
Retrotransposons	LINE/CR1	L3	ABGB01001299	171	240
Retrotransposons	LINE/CR1	L3	ABGB01001329	79	148
Retrotransposons	LINE/CR1	L3	ABGB01001419	72	141
Retrotransposons	LINE/CR1	L3	ABGB01001542	362	431
Retrotransposons	LINE/L2	L2d	CH991542	78241	78311
Retrotransposons	LINE/Penelope	Penelope1_Vert	ABGB01000020	49017	49051
Retrotransposons	LINE/RTE-BovB	MamRTE1	ABGB01000017	75292	75340
Retrotransposons	SINE/MIR	MIRc	ABGB01000037	7063	7121
Retrotransposons	SINE/tRNA-RTE	MamSINE1	ABGB01000213	749	780

Retrotransposons	SINE/tRNA-RTE	MamSINE1	ABGB01000234	904	935
Retrotransposons	SINE/tRNA-RTE	MamSINE1	ABGB01000465	44	75
Retrotransposons	SINE/tRNA-RTE	MamSINE1	ABGB01000922	320	351
Retrotransposons	SINE/tRNA-RTE	MamSINE1	ABGB01001291	330	361
Retrotransposons	SINE/tRNA-RTE	MamSINE1	ABGB01001299	469	500
Retrotransposons	SINE/tRNA-RTE	MamSINE1	ABGB01001419	370	401
Retrotransposons	SINE/tRNA-RTE	MamSINE1	ABGB01001542	660	691
Retrotransposons	SINE/tRNA-RTE	MamSINE1	CH991540	153301	153365
Transposon	DNA/hAT-Ac	OldhAT1	ABGB01000117	1000	1043
Transposon	DNA/hAT-Ac	OldhAT1	ABGB01000183	400	455
Transposon	DNA/hAT-Ac	OldhAT1	ABGB01000373	991	1046
Transposon	DNA/hAT-Ac	OldhAT1	ABGB01000523	688	731
Transposon	DNA/hAT-Ac	OldhAT1	ABGB01000598	1016	1060
Transposon	DNA/hAT-Ac	OldhAT1	ABGB01000718	982	1025
Transposon	DNA/hAT-Ac	OldhAT1	ABGB01000868	1033	1076
Transposon	DNA/hAT-Ac	OldhAT1	ABGB01001115	794	837
Transposon	DNA/hAT-Ac	OldhAT1	ABGB01001163	317	360
Transposon	DNA/hAT-Ac	OldhAT1	ABGB01001524	316	359
Transposon	DNA/hAT-Charlie	MER58D	ABGB01000014	31157	31224
Transposon	DNA/PIF-Harbinger	Chompy-6_Croc	CH991541	33532	33595
Transposon	DNA/TcMar-Tigger	Tigger19b	ABGB01000027	16712	16810
Transposon	DNA/TcMar-Tigger	Tigger19a	ABGB01000243	757	813

Table 7: REPEATFINDER transposable element predictions for the genome of *E. hepatopenaei*

Type	class/family	repeat	query	begin	end
Retrotransposon	LINE/CR1	X8_LINE	lcl EHP_24	13519	13550
Retrotransposon	LINE/CR1	L3	lcl EHP_277	3619	3681
Retrotransposon	LINE/CR1	X6B_LINE	lcl EHP_80	3234	3312
Retrotransposon	LINE/CR1	L3b	lcl EHP_83	3179	3222
Retrotransposon	LINE/L1	L1M3	lcl EHP_3872	87	332
Retrotransposon	LINE/L1	L1M3	lcl EHP_830	1839	2084
Retrotransposon	LINE/L2	AmnL2-1	lcl EHP_10	41522	41584
Retrotransposon	LINE/L2	AmnL2-1	lcl EHP_4	45249	45293
Retrotransposon	LINE/L2	AmnL2-1	lcl EHP_40	29397	29444
Retrotransposon	LINE/RTE-X	L4_C_Mam	lcl EHP_3872	288	346
Retrotransposon	LINE/RTE-X	L4_C_Mam	lcl EHP_830	2040	2098
Retrotransposon	SINE/MIR	MIRc	lcl EHP_83	20506	20548
Retrotransposon	SINE/tRNA-RTE	MamSINE1	lcl EHP_223	7028	7072

Table 8: REPEATFINDER transposable element predictions for the genome of *H. eriocheir canceri*

Type	class/family	repeat	query	begin	end
Retrotransposon	LINE/L2	L2	lcl scaffold_1207	366	427
Retrotransposon	LINE/L2	L2	lcl scaffold_227	283	344
Retrotransposon	LINE/L2	L2	lcl scaffold_2598	584	645
Retrotransposon	LINE/L2	L2d	lcl scaffold_47	1997	2028
Retrotransposon	LINE/CR1	L3	lcl scaffold_1	5936	5973
Retrotransposon	LINE/CR1	L3	lcl scaffold_433	1584	1617
Retrotransposon	SINE/MIR	MIR	lcl scaffold_88	3608	3663

Retrotransposon	LINE/CR1	X17_LINE	lcl scaffold_129	7649	7693
Retrotransposon	LINE/CR1	X8_LIINE	lcl scaffold_225	2850	2925
Transposon	DNA/Merlin	DNA/Merlin	lcl scaffold_381	408	501

Table 9: REPEATFINDER transposable element predictions for the genome of *H. eriocheir*

Type	class/family	repeat	query	begin	end
Retrotransposon	LINE/CR1	L3	NODE_3632	41361	41398
Retrotransposon	LINE/CR1	L3	NODE_744	3226	3259
Retrotransposon	LINE/CR1	X17_LINE	NODE_8586	8052	8096
Retrotransposon	LINE/L1	X9_LINE	NODE_6192	6465	6594
Retrotransposon	LINE/L2	AmnL2-1	NODE_1101	1314	1369
Retrotransposon	LINE/L2	L2d	NODE_5603	131	162
Retrotransposon	LINE/L2	L2	NODE_7395	26660	26721
Retrotransposon	LINE/L2	L2d	NODE_7849	25861	25892
Retrotransposon	LINE/RTE-BovB	MamRTE1	NODE_19	30937	31037
Retrotransposon	LINE/RTE-BovB	MamRTE1	NODE_2943	718	820
Retrotransposon	SINE/MIR	MIR	NODE_6100	5321	5376
Transpososn	DNA/hAT-Charlie	Charlie2b	NODE_5812	40801	40845
Transpososn	DNA/hAT-Charlie	Charlie17	NODE_980	3293	3394
Transpososn	DNA/Merlin	Merlin1_HS	NODE_496	312	405

Table 10: REPEATFINDER transposable element predictions for the genome of *Enc. canceri*

Type	class/family	repeat	query	begin	end
Retrotransposon	LINE/CR1	CR1-L3A_Croc	scaffold_1191	583	767
Retrotransposon	LINE/CR1	CR1_Mam	scaffold_121	2878	2932
Retrotransposon	LINE/CR1	X2_LINE	scaffold_121	3065	4258
Retrotransposon	LINE/CR1	X2_LINE	scaffold_123	3813	4191
Retrotransposon	LINE/CR1	CR1_Mam	scaffold_1291	33	105
Retrotransposon	LINE/CR1	Plat_L3	scaffold_1291	72	700
Retrotransposon	LINE/CR1	Plat_L3	scaffold_1475	194	665
Retrotransposon	LINE/CR1	X2_LINE	scaffold_149	1	1225
Retrotransposon	LINE/CR1	CR1_Mam	scaffold_1592	131	626
Retrotransposon	LINE/CR1	Plat_L3	scaffold_1692	2	437
Retrotransposon	LINE/CR1	CR1_Mam	scaffold_172	639	1736
Retrotransposon	LINE/CR1	X2_LINE	scaffold_276	6	324
Retrotransposon	LINE/CR1	L3	scaffold_29	9907	9970
Retrotransposon	LINE/CR1	X2_LINE	scaffold_294	160	549
Retrotransposon	LINE/CR1	CR1_Mam	scaffold_490	580	1256
Retrotransposon	LINE/CR1	CR1-1_Amn	scaffold_493	1	415
Retrotransposon	LINE/CR1	CR1_Mam	scaffold_585	228	1059
Retrotransposon	LINE/CR1	CR1-3_Croc	scaffold_604	39	117
Retrotransposon	LINE/CR1	X2_LINE	scaffold_698	13	391
Retrotransposon	LINE/CR1	CR1-1_Amn	scaffold_714	18	861
Retrotransposon	LINE/CR1	CR1-L3A_Croc	scaffold_888	1	295
Retrotransposon	LINE/CR1	X2_LINE	scaffold_922	227	869
Retrotransposon	LINE/CR1	X2_LINE	scaffold_976	4	858
Retrotransposon	LINE/CR1	Plat_L3	scaffold_997	6	554
Retrotransposon	LINE/L1	L1M3f	scaffold_34	23390	23432

Retrotransposon	LINE/L2	X13_LINE	scaffold_25	16983	17040
Retrotransposon	LINE/RTE-BovB	MamRTE2	scaffold_2	29891	30014
Retrotransposon	LTR/ERV1	LTR12D	scaffold_10	28073	28157
Retrotransposon	LTR/Gypsy	MamGypLTR2b	scaffold_2	67809	67876
Retrotransposon	LTR/Gypsy	MamGypsy2-l	scaffold_99	3442	3527
Transposon	DNA/hAT-Charlie	Charlie6	scaffold_472	165	240
Transposon	DNA/hAT-Charlie	Charlie8	scaffold_5	2032	2150

Appendix 10: Plasma membrane transporter predictions

Table 1: Comparing plasma membrane transporter complement between members and non-members of the Enterocytozoonidae family.

TARGET SUBSTRATE	PROTEIN FAMILY/DOMAIN	<i>E. bienersi</i>	<i>Ent. canceri</i>	<i>E. hepatoparvae</i>	<i>H. eriocheir</i>	<i>H. eriocheir canceri</i>	<i>Enc. cuticuli</i>	<i>T. hominis</i>
MULTI-SUBSTRATE	MFS	0	3	3	3	0	0	3
	ABC	2	7	3	3	1	1	8
	Sulp	1	1	0	0	1	1	1
	MIP	1	2	1	1	1	1	1
	P-typeATPases	3	4	3	1	4	4	7
LIPIDS	CTL	0	3	1	1	0	1	1
	Di-sulfide bridge nucleocytoplasmic transport domain	0	0	0	0	0	0	0
PROTEINS/POLYPEPTIDES	POT	0	0	0	0	0	0	1
	Transmembrane amino acid transporter	0	3	2	2	0	0	5
AMINO ACIDS	Amino Acid Permease	1	1	1	1	0	1	1
	V/F-type ATPases	1	3	3	3	2	2	4
NUCLEOTIDES	AAA/TLC ATP-ADP transporters	3	4	4	4	3	3	4
	UAA	3	3	1	1	1	1	2
	DMT	0	1	1	1	1	1	1
	MScS	3	2	2	2	0	0	5
	Cation proton antiporter CPA1	0	1	1	1	0	0	1
IONS	DASS	1	0	0	0	0	0	0
	Ca-ClC	0	1	0	0	1	1	1
	Cation efflux family	1	2	0	0	0	0	1
	ZIP	1	1	1	1	1	1	1
	HCO ₃ transporter	0	0	0	0	0	0	1
SUGAR	Glucose transporter/Sugar permease	1	2	2	2	0	0	2
	PHO1	1	1	1	1	1	1	1
INORGANIC PHOSPHATE	Total number of predicted plasma membrane transporters assigned to a protein family	23	45	30	17	17	17	48
	Paralogous unique to species	4	0	6	0	0	0	0
DISTRIBUTION OF PREDICTED TRANSPORTERS NOT ASSIGNED TO A PROTEIN FAMILY	Orthologous but only found within the Enterocytozoonidae	2	5	11	2	1	0	0
	Non-orthologous	10	0	1	0	0	0	1
	Orthologs found in Enterocytozoonidae and in non Enterocytozoonidae	128	105	103	92	82	111	107
Total number of predicted plasma membrane transporters not assigned to a protein family		144	110	121	94	83	111	108
TOTAL NUMBER OF PREDICTED PLASMA MEMBRANE TRANSPORTERS		167	155	151	111	100	159	154

Appendix 11: Taxonomic profile of *Enterospora canceri*'s RUN1 assembly

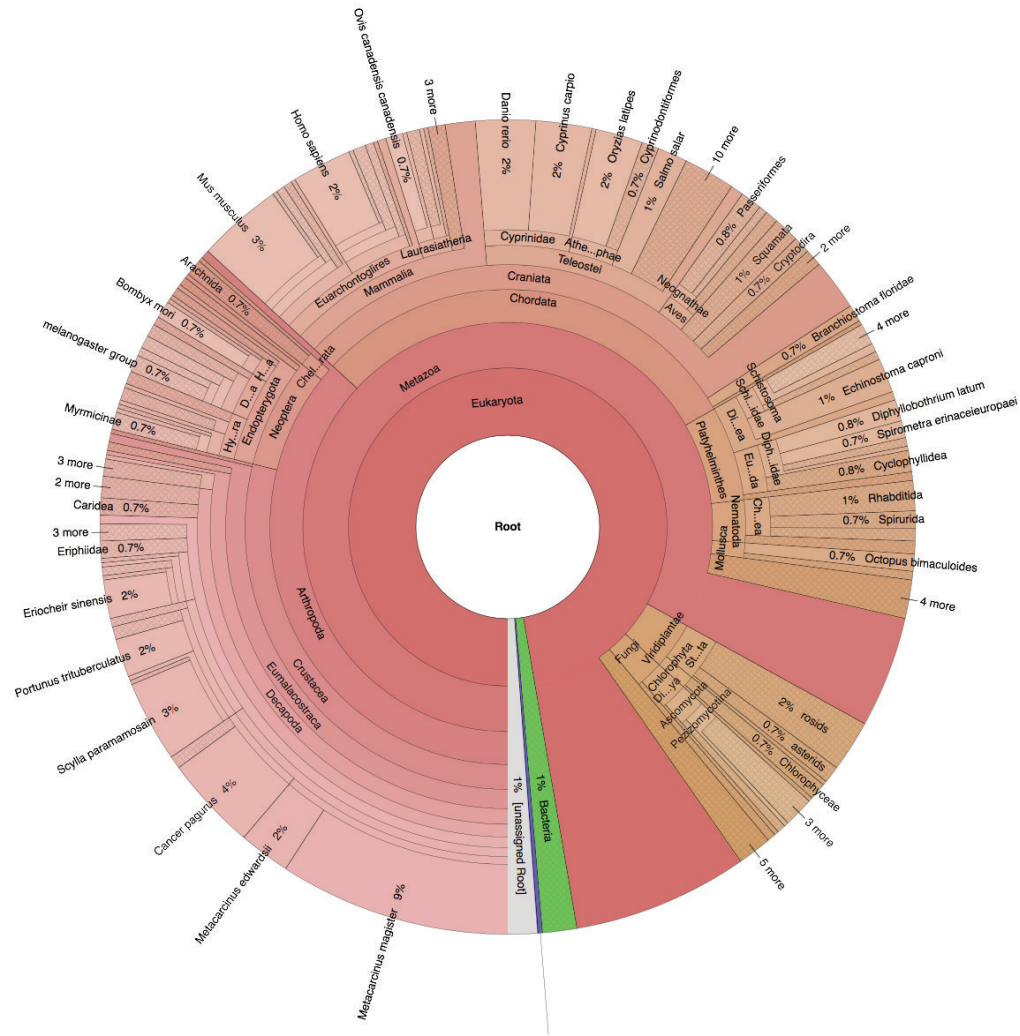


Figure 1: Taxonomic profile of *Enterospora canceri*'s unfiltered RUN1 contigs

Appendix 12: Bateman *et al.*, 2016 Paper published in the Parasitology Journal

971

Single and multi-gene phylogeny of *Hepatospora* (Microsporidia) – a generalist pathogen of farmed and wild crustacean hosts

K. S. BATEMAN^{1*}, D. WIREDU-BOAKYE², R. KERR¹, B. A. P. WILLIAMS² and G. D. STENTIFORD¹

¹ European Union Reference Laboratory for Crustacean Diseases, Centre for Environment, Fisheries and Aquaculture Science (Cefas), Weymouth Laboratory, Weymouth, Dorset, DT4 8UB, UK

² Biosciences, University of Exeter, Geoffrey Pope Building, Stocker Road, Devon, EX4 4EU, UK

(Received 3 December 2015; revised 12 February 2016; accepted 23 February 2016; first published online 22 March 2016)

SUMMARY

Almost half of all known microsporidian taxa infect aquatic animals. Of these, many cause disease in arthropods. *Hepatospora*, a recently erected genus, infects epithelial cells of the hepatopancreas of wild and farmed decapod crustaceans. We isolated *Hepatospora* spp. from three different crustacean hosts, inhabiting different habitats and niches; marine edible crab (*Cancer pagurus*), estuarine and freshwater Chinese mitten crab (*Eriocheir sinensis*) and the marine mussel symbiont pea crab (*Pinnotheres pisum*). Isolates were initially compared using histology and electron microscopy revealing variation in size, polar filament arrangement and nuclear development. However, sequence analysis of the partial SSU rDNA gene could not distinguish between the isolates (~99% similarity). In an attempt to resolve the relationship between *Hepatospora* isolated from *E. sinensis* and *C. pagurus*, six additional gene sequences were mined from on-going unpublished genome projects (RNA polymerase, arginyl tRNA synthetase, prolyl tRNA synthetase, chitin synthase, beta tubulin and heat shock protein 70). Primers were designed based on the above gene sequences to analyse *Hepatospora* isolated from pea crab. Despite application of gene sequences to concatenated phylogenies, we were unable to discriminate *Hepatospora* isolates obtained from these hosts and concluded that they likely represent a single species or, at least subspecies thereof. In this instance, concatenated phylogenetic analysis supported the SSU-based phylogeny, and further, demonstrated that microsporidian taxonomies based upon morphology alone are unreliable, even at the level of the species. Our data, together with description of *H. eriocheir* in Asian crab farms, reveal a preponderance for microvariants of this parasite to infect the gut of a wide array of decapods crustacean hosts and the potential for *Hepatospora* to exist as a cline across wide geographies and habitats.

Key words: edible crab, pea crab, Chinese mitten crab, microsporidian, *Hepatospora*, multi-gene phylogeny, taxonomy, Enterocytozoonidae.

INTRODUCTION

Microsporidia are single-celled eukaryotic intracellular parasites known to infect a range of vertebrate and invertebrate hosts (Mathis *et al.* 2005). Since the inception of the phylum 'Microsporidia', both phylogenetic placement of the group and classification within the group have proven problematic [reviewed in (Corradi and Keeling, 2009)]. Early phylogenies used to place microsporidia within the tree of life failed to account for rate heterogeneity among gene sites, base-compositional biases and the overall accelerated evolutionary rate characteristic of microsporidian genomes (Hirt *et al.* 1999) and thus microsporidia were for a long-time considered basal eukaryotes. They are now, along with the Cryptomycota, considered the most basal fungi

group, a new classification based on the phylogenetic analysis of 200 genes (James *et al.* 2013).

Within the phylum, in the past, classification of taxa has been based on structural characteristics, ultrastructural morphology and karyotypic evidence (Canning, 1953; Vavra and Undeen, 1970; Shaddock *et al.* 1990; Cali *et al.* 1993). Now there is a compounding body of evidence supporting the idea that morphological and developmental features in this phylum are plastic between both closely and distantly related microsporidia (Vossbrinck and Debrunner-Vossbrinck, 2005; Stentiford *et al.* 2013). Stentiford *et al.* (2013) observed that microsporidians isolated from marine decapod crustaceans that would have been classed as distantly related taxa (*Nadelspora* and *Ameson*) under a morphology-based classification system are in fact close relatives on rDNA-based phylogenetic trees and are potential life-cycle variants of the same taxon. An example of a highly plastic morphological character used in classification is nuclear configuration. Two configurations are known to exist in microsporidia: a

* Corresponding author: European Union Reference Laboratory for Crustacean Diseases, Centre for Environment, Fisheries and Aquaculture Science (Cefas), Weymouth Laboratory, Weymouth, Dorset DT4 8UB, UK. E-mail: kelly.bateman@cefasc.co.uk

monokaryon (an individual nucleus) and a diplokaryon (two apposed nuclei). Whilst some species retain the same nuclear configuration throughout their life cycle, others switch between the nuclear stages depending upon life-cycle stage, and polymorphic species shift between the nuclear stages when changing hosts or tissues (Vávra and Larsson, 2014).

For many years the alternative to morphological characters has been the SSU rRNA gene, which has now been amplified and sequenced from over 1000 species of microsporidia. However, the rDNA sequences amplified are generally only able to resolve taxonomic relationships down to the genus level (Vossbrinck *et al.* 1998; Vossbrinck and Debrunner-Vossbrinck, 2005). The most commonly used primers, F18 and 1492R (Vossbrinck *et al.* 1993; Kent *et al.* 1996) amplify a 1400 bp fragment of the most conserved region of the SSU rDNA and omit the internal transcribed spacer (ITS) (Vossbrinck and Debrunner-Vossbrinck, 2005), a region reported to be highly variable even within species and hence highly informative for intra-species differentiation analyses (Gresoviac *et al.* 2000; Sak *et al.* 2011). However, even if this region was commonly amplified for microsporidia, there are other issues with using SSU as a marker for closely related species. In at least some microsporidia, the process of concerted evolution that typically keeps rDNA copies uniform within the genome, does not seem to be present. For example, Ironside (2013), O'Mahony *et al.* (2007) and Tay *et al.* (2005) observed a higher ITS sequence variation between repeats in the same genome than between different genomes for some *Nosema* species, which means that caution needs to be used when employing rDNA to discriminate between closely related microsporidia.

One of the issues emerging from the difficulty in resolving close phylogenetic relationships and morphological plasticity in microsporidians is the assignment of appropriate species names. This is an important issue because taxonomic names of pathogenic species are fed into legislative frameworks that are used to inform policy making (Stentiford *et al.* 2014). Therefore there is a growing need to find alternative molecular markers to resolve questions in microsporidia taxonomy. There is currently whole-genome data for 20 microsporidian species on the publicly available MicrosporidiaDB database. With cheaper thorough-put sequencing technologies and the advent of single cell genome sequencing, we can only expect this number to increase. This provides a minable resource for molecular characters for multi-locus phylogenies (Capella-gutiérrez *et al.* 2012).

We isolated putative *Hepatospora* sp. parasites from three different crustacean hosts, inhabiting different habitats and niches; marine edible crab (*Cancer pagurus*), estuarine and freshwater Chinese

mitten crab (*Eriocheir sinensis*) and the marine mussel symbiont pea crab (*Pinnotheres pisum*). Isolates were initially compared using histology and electron microscopy revealing not only a similar life cycle, but also the variation in size, polar filament arrangement and nuclear development (Table 1). However, sequence analysis of the partial SSU rDNA gene could not distinguish between the isolates ($\approx 99\%$ similarity). Here we take advantage of the current microsporidian whole-genome database and our in-house genomic data to construct a six-gene concatenated phylogenetic tree for *Hepatospora eriocheir*, a parasite of the invasive Chinese mitten crab (Stentiford *et al.* 2011), a pea crab (*P. pisum*) infecting microsporidium (Longshaw *et al.* 2012) and a novel microsporidium that infects commercially important edible crabs (*C. pagurus*), all from European waters. Our results revealed that *Hepatospora* isolates from the three different crustacean hosts are likely to be microvariants of a single species. Further, they support the concept that microsporidian taxonomies based upon morphology are not only unreliable, but can also be deceptive, even at the level of the species. Our data, together with description of *H. eriocheir* as an agent of emergent disease in Asian crab aquaculture (Ding *et al.* 2016) reveal a preponderance for microvariants of *H. eriocheir* to infect the gut of a wide array of decapods crustacean hosts across wide geographic boundaries and in a range of habitats.

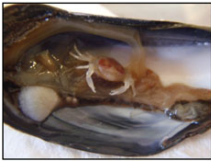


MATERIALS AND METHODS

Eriocheir sinensis, *Cancer pagurus* and *Pinnotheres pisum* sampling

Chinese mitten crabs (*E. sinensis*) were sampled from two locations in the Thames Estuary; a site near to the Millennium Dome (51:27:12N, 00-00-44E) and another at Tilbury Power Station (51:27:12N, 00-23-10E). Crabs were collected using Fyke nets and from the screens of water intake pipes at the power station. Edible crabs (*C. pagurus*) were captured using baited pots in the Weymouth and Portland area of the English Channel, UK (50°32'50"N, 002°11'00"W) as previously described (Bateman *et al.* 2011). Live crabs were transported to the Cefas laboratory in Weymouth and anaesthetized by chilling on ice for 30 min before dissection. The hepatopancreas, gill, gonad, central nerve ganglia, heart and body muscle was removed from the crabs and fixed in Davidson's sea water fixative for histology. Additional hepatopancreas samples were fixed in 2.5% glutaraldehyde in 0.1 M sodium cacodylate buffer for electron microscopy, and 100% ethanol and frozen, for molecular analyses.

For *P. pisum* sampling, blue mussels (*Mytilus edulis*) were collected from a range of sites around the UK. Mussels were opened by cutting the

Table 1. Comparison of host and structure [histological, electron microscopy and molecular data (SSU 18S rRNA)] from the *Hepatospora* isolates

			
Host			
Species	<i>Pinnotheres pisum</i>	<i>Eriocheir sinensis</i>	<i>Cancer pagurus</i>
Family	Pinnotheridae	Varunidae	Cancridae
Habitat	Symbiont of marine mussels	Estuarine	Marine
Microsporidian			
Species	<i>Hepatospora</i> sp.	<i>Hepatospora eriocheir</i>	<i>Hepatospora</i> sp.
Pathology			
Tissue	Hepatopancreas	Hepatopancreas	Hepatopancreas
Site of development	Cytoplasm of hepatopancreatic epithelial cells	Cytoplasm of hepatopancreatic epithelial cells	Cytoplasm of hepatopancreatic epithelial cells
Ultrastructure			
Spore	Ellipsoid	Ellipsoid	Ellipsoid
Size	1.9 × 0.9 μm	1.8 × 0.9 μm	1.8 × 0.9 μm
Polar filament	Isofilar	Isofilar	Isofilar
Polar filament turns	5–6 polar filament coils in a single rank	7–8 polar filament coils in a single rank	7–8 polar filament coils in a single rank
Nuclear status of spore	Diplokaryotic	Unikaryotic	Diplokaryotic
Development	Syncronous development within a parasitophorous vacuole	Syncronous development within a parasitophorous vacuole	Syncronous development within a parasitophorous vacuole
Molecular analysis			
GenBank Accession	–	HE584635.1	HE584633.1
SSU similarity between isolates	99%	100%	100%

abductor muscle and separating the valves. Symbiotic pea crabs (*P. pisum*) found habituating the mantle of the host mussel were removed and anaesthetized on ice prior to bilateral dissection and removal of the hepatopancreas. Small portions of the organ were fixed in Davidson's sea water fixative, 2.5% glutaraldehyde in 0.1 M sodium cacodylate buffer and 100% ethanol for histology, electron microscopy and molecular techniques, respectively.

Histology and transmission electron microscopy

For histology, fixation was allowed to proceed for 24 h before samples were transferred to 70% industrial methylated spirit. Fixed samples were processed to wax in a vacuum infiltration processor using standard protocols. Sections were cut at a thickness of 3–5 μm on a rotary microtome and mounted onto glass slides before staining with haematoxylin and eosin (H&E) and Feulgen stains. Stained sections were analysed by light microscopy (Nikon Eclipse E800) and digital images and measurements were taken using the Lucia™ Screen Measurement System (Nikon, UK). For electron microscopy, tissues were fixed in 2.5% glutaraldehyde in 0.1 M

sodium cacodylate buffer (pH 7.4) for 2 h at room temperature and rinsed in 0.1 M sodium cacodylate buffer (pH 7.4). Tissues were post-fixed for 1 h in 1% osmium tetroxide in 0.1 M sodium cacodylate buffer. Samples were washed in three changes of 0.1 M sodium cacodylate buffer before dehydration through a graded acetone series. Samples were embedded in Agar 100 epoxy (Agar Scientific, Agar 100 pre-mix kit medium) and polymerized overnight at 60 °C in an oven. Semi-thin (1–2 μm) sections were stained with Toluidine Blue for viewing with a light microscope to identify suitable target areas. Ultrathin sections (70–90 nm) of these areas were mounted on uncoated copper grids and stained with 2% aqueous uranyl acetate and Reynolds' lead citrate (Reynolds, 1963). Grids were examined using a JEOL JEM 1400 transmission electron microscope and digital images captured using an AMT XR80 camera and AMTv602 software.

Spore isolation and DNA extraction

Hepatopancreas samples from histologically confirmed microsporidian-infected crabs were crushed with a sterile pestle and mortar in PBS. The homogenous mash was filtered through a 100

μm mesh followed by cell sieving through 40 μm filter. Filtrate was topped up to 50 mL using PBS/triton-X (0.1%) and pelleted at 3220 g in an Eppendorf centrifuge precooled at 4 °C, for 10 min. The supernatant was removed and the pellet (containing host cell debris and microsporidian spores) was resuspended in 2.5 mL of ice-cold water. Homogenate was added to the top of a Percoll density gradient and centrifuged at 1000 g in a pre-cooled centrifuge for 45 min. Phase contrast and fluorescence microscopy were used to visualize the spores and to examine for purity following Percoll gradient purification. Spores were stained with the chitin marker calcofluor white (Darken, 1962).

For extraction of DNA, spores of the Pea crab and Mitten crab parasites were diluted 1 in 10 (w/v) in G2 buffer and Proteinase K was added at a concentration of 2 mg mL⁻¹ (Qiagen, UK). The spores were subsequently disrupted in a Matrix D FastPrep cell disrupter (FastPrep, UK) by shaking on a homogenizer for 2 min at highest setting. Homogenized samples were incubated for 4 h at 56 °C. Total DNA was extracted using an EZ1 DNA tissue kit and EZ1 Advanced XL BioRobot (Qiagen) following manufacturers' instructions. For the edible crab parasite, aliquots of purified spores suspended in 1 × PBS were mixed with liquid nitrogen in a sterile mortar. The mixture was slowly stirred until it solidified, after which it was ground with a sterile pestle for 10 min. Liquid nitrogen was again added to the powder and the mix was ground for an additional 10 min. This step was repeated three times before dissolving the resulting powder in 800 μL of phenol (pH 8.0). The homogenate was transferred to an Eppendorf tube and mixed by inversion and subsequently centrifuged for 10 min at 10 000 × g. The recovered supernatant was mixed by inversion with 400 μL of chloroform and centrifuged for another 10 min at 10 000 × g. Genomic DNA was precipitated from the aqueous solution using a standard ethanol precipitation protocol (Ausubel *et al.* 2002) and sent to the University of Exeter sequencing service, UK for library preparation and Illumina sequencing.

Identification of six marker genes from unpublished genome projects of H. eriocheir and the edible crab microsporidian

Orthologues of the six marker genes [amino acyl tRNA synthetases (Brown and Doolittle, 1999); arginyl tRNA synthetase and prolyl tRNA synthetase, beta-tubulin (Edlind *et al.* 1994), chitin synthase (Hinkle *et al.* 1997), heat shock protein 70 (HSP70) (Hirt *et al.* 1997) and RNA polymerase II (Hirt *et al.* 1999)] for publicly available microsporidian genomes were obtained from the MicrosporidiaDB database (Aurrecochea *et al.* 2011) by initially

performing word searches for the individual marker gene names for *Encephalitozoon cuniculi* GB-M1 and then using the 'transform by orthology tool' to find orthologues for all the other microsporidians in the database. To identify the desired marker genes in the newly sequenced genomes of *H. eriocheir* and the edible crab microsporidian, predicted open reading frames (ORFs) were queried using the microsporidian proteins obtained from MicrosporidiaDB (Aurrecochea *et al.* 2011) using command line blastn (Mount, 2007) with an e-value cutoff of 1e-5. The top ORF hits were selected as the orthologous genes and orthology was verified by assessment of phylogenies of single genes.

Primer design, PCR and sequencing of the six-marker gene from the pea crab parasite

Due to the small amount of pea crab parasite genomic DNA recovered from the extraction procedure, gene-specific PCRs and subsequent sequencing was performed to retrieve the corresponding sequences for the six genes of the pea crab parasite rather than full genome sequencing. We designed gene-specific primers by using the first and last 18 nucleotides of the selected orthologues from *H. eriocheir*. A two-step nested PCR was done to amplify longer genes such as Arginyl tRNA and RNA polymerase (Table 2).

Ribosomal DNA phylogenetic analyses

Universal primers were used to amplify rDNA regions from genomic material extracted from *H. eriocheir*, pea crab and edible crab parasites (Table 3). The sequencing results for the PCR products were subsequently aligned and used to construct a neighbour-joining tree with MEGA software (Tamura *et al.* 2011). All PCR reactions were performed in a 50 μL reaction mix consisting of 1 × Green Go Taq buffer, 2.5 mM MgCl₂, 0.25 mM dNTPs, 100 pM each of the forward and reverse primer sets, 0.25 units Go Taq Flexi (Promega, UK) and 2.5 μL of extracted genomic DNA. Amplifications were performed on a Peltier PTC-225 thermal cycler with the following settings: Initialization step at 94 °C for 5 min, 40 cycles of 1 min denaturation at 95 °C, a 1 min annealing step (see Tables 1 and 2 for temperatures) and a 1 min extension step at 72 °C. A final elongation step was carried out for 10 min at 72 °C. Amplification products were resolved on 2% agarose gels stained with ethidium bromide and visualized using a UV illuminator.

Correct size products (Tables 2 and 3) were excised from the gels and purified using the Wizard SV gel and PCR purification system (Promega, UK). PCR products were sequenced using Sanger technology, ABI PRISM Big Dye

Table 2. Gene-specific primers were designed using the first and last 18 nucleotides of the selected orthologues from *H. eriocheir*

Gene	Primers second round	Sequence	Annealing temp (°C)	Size of product (bp in Mitten crab)
Beta tubulin first round	F1 R1(R&C)	GTAAGTGATACAGTTGTAGAACC CCTTCACCAGTGTACCAGTG	55	683
Beta tubulin second round	F1 R2(R&C)	GTAAGTGATACAGTTGTAGAACC CATTATTAGGAATCCACTCAAC	55	523
	F2 R1(R&C)	GTTGAGTGGATTCTTAATAATG CCTTCACCAGTGTACCAGTG	55	182
Prolyl tRNA first round	F1 R1(R&C)	ATGAAGATTTATTAGCTGTGCC GGAATACCTTTAAGTTCGCAG	55	496
Prolyl tRNA second round	F1 R3(R&C)	ATGAAGATTTATTAGCTGTGCC CTCAGAAGCTACGTCAAC	55	200
	F2 R1(R&C)	GTGATGACGTAGCTTCTGAG GGAATACCTTTAAGTTCGCAG	55	316
Arginyl tRNA first round	F1 R1(R&C)	ATGGAAGAAGGTATAAATGAGG GGTCCCAGTGATTCACCTTT	55	668
Arginyl tRNA first & second round	F2 R1(R&C)	TTAGTTACAGGCATGTCAACC GGTCCCAGTGATTCACCTTT	55	242
	F1 R2(R&C)	ATGGAAGAAGGTATAAATGAGG GGTTGACATGCCGTAACTAA	55	447
HSP70 first round	F1 R2(R&C)	AGAGACAAGCAACAAAAGATGC CAGCTGATATTTTCAACTTATTCAA	57	336
HSP70 first & second round	F1 R1(R&C)	AGAGACAAGCAACAAAAGATGC CACTATCAAAATCTTCTCCTCC	57	240
	F2 R2(R&C)	GGAGGAGAAGATTTTGATAGTG CAGCTGATATTTTCAACTTATTCAA	57	118
RNA polymerase first round	F1 R1(R&C)	AGTGAGATTAGATCTGTACCTG GGATGTATTTTCACAATGTGTATAT	57	1400
RNA polymerase first & second round	F1 R2(R&C)	AGTGAGATTAGATCTGTACCTG GGTGTATTTACACGCTTAAATG	57	779
	F2 R1(R&C)	CATTTAAGACGTGTAATACACC GGATGTATTTTCACAATGTGTATAT	57	644
Chitin synthase first round	F1 R1(R&C)	TGACAGGATGAGTGATGTGG GACTAATATAACTCAAACACTT	55	254

Terminator v3.1 cycle sequencing kit (Thermo Fisher Scientific, UK) with relevant primers (Tables 2 and 3) and the following PCR thermal cycling program: 94 °C × 30 s followed by 30 cycles of 96 °C × 10 s, 50 °C × 10 s and 60 °C × 4 min and held at 4 °C. Analyses of the sequenced PCR product were done using the Sequencher software (Gene codes corporation).

Construction of a six-gene phylogenetic tree

Maximum likelihood (ML) analyses. The orthologues for each gene set were individually aligned with the command line MUSCLE program (v3.8.31) (Edgar, 2004) using the default settings, and then subsequently masked with the automatic command line tool TrimAl (Capella-Gutierrez *et al.* 2009). *Saccharomyces cerevisiae* was used as an out-group in each of the microsporidian datasets.

A GTR substitution model with a gamma model of rate heterogeneity was used to create ML trees for the individual gene sets with the RaxML program (Stamatakis, 2014). These were pilot trees to check for unusually long-branch lengths

indicative of unlikely orthologues. The masked genes from each microsporidian species were subsequently manually concatenated using SeaView (v4) (Gouy *et al.* 2010). For the final construction of the ML concatenated gene tree, a partition file that contained the positions of the individual genes within the alignment was manually created and passed to the RaxML program using the '-q' option (Stamatakis, 2014). This was to enable the program to treat each gene set in the concatenated alignment separately and allow it to estimate individual nucleotide substitution rates. These estimations were also performed with the GTR + GAMMA nucleotide substitution model.

Bayesian inference analysis on six-gene concatenated alignment. To check for reliability of the phylogenetic relationships estimated by ML analyses, a Bayesian inference method was also used to reconstruct the six-gene concatenated phylogenetic tree using MrBayes program (v3.2) (Ronquist *et al.* 2012). A partition file containing positions of the individual genes in the alignment was created according to the program manual. The program was run

Table 3. rDNA primers used in this study

Gene	Primer name	Sequence	Annealing temp (°C)	Size of product approx. bp	Reference
SSU	530 F	GTGCCATCCAGCCGGG	55	1350–1550	Docker <i>et al.</i> (1997)
	580 R	GGTCCGTGTTTCAAGACGG			
SSU	MF1	CCGGAGAGGGAGCCTGAGA	55	848	Tourtip <i>et al.</i> (2009)
	MR1	GACGGGCGGTGTGTACAAA			
SSU	Medlin B (used with MF1)	GATCC'TTCTGCAGG'TTACCT'	55	1500	Medlin <i>et al.</i> (1988)

using a GTR + GAMMA model and probability distributions were generated using the Markov Chain Monte Carlo Methods. A total of 1 020 000 generations were run, the first 25% of sampled trees were discarded as 'burn-in' and a consensus tree was constructed.

RESULTS

Pathology, ultrastructure and SSU-based phylogeny

Infected hepatopancreatic tubules from all three crab hosts displayed varying proportions of infected epithelial cells consistent with previous descriptions of *Hepatospora*-associated pathology (Stentiford *et al.* 2011). Infection was confined to the epithelial cells of the hepatopancreas, with no other organ seemingly infected. However, in some severe cases where hepatopancreatic tubules were disrupted, liberated parasite spores could be observed within the lumen of affected tubules and, in the haemolymph. Histologically, the disease caused by this microsporidian was indistinguishable between Chinese mitten crab (where the parasite has been confirmed as *H. eriocheir*), edible crab and pea crab (Fig. 1).

Ultrastructural analysis revealed multiple stages of a microsporidian parasite within the cytoplasm of hepatopancreatic epithelial cells of each host crab species. The earliest stage observed was the meront (Fig 2A, C and E). *Hepatospora eriocheir* (infecting Chinese mitten crab) possessed uni-nucleate meronts, whereas the parasite infecting pea crab and edible crab possessed bi-nucleate meronts. A similar observation was made in the spore stages with spores (Fig 2B, D and F) from the parasite infecting edible crab and pea crab appearing bi-nucleate, while spores from the Chinese mitten crab were of uni-nucleate karyotype. Mature spores in all cases possessed a trilaminar wall consisting of a plasma membrane, an electron lucent endospore and an electron-dense exospore. Spores measured 1.8–1.9 × 0.9 µm but contained varying turns of an isofilar polar filament. Spores from the parasite-infecting Chinese mitten crabs and edible crabs possessed seven to eight turns of the polar filament while those from the parasite-infecting pea crabs possessed five to six turns of the polar filament. A summary of these shared and distinctive features is given in Table 1.

Partial sequencing of the SSU rDNA gene obtained from the parasites infecting the three host crab species revealed an apparent synonymy (≈99% similarity over 890 bp), at least based upon this portion of the SSU, between the three parasites. Despite some ultrastructural and karyotypic distinctions between the three isolates (Table 1), SSU-based phylogeny did not support erection of distinctive taxa for the parasites infecting edible crab and pea crab. Based upon SSU phylogeny, these parasites would therefore be classified as *H. eriocheir*, with edible crab and pea crab representing an extended host range for this parasite.

Phylogeny of *Hepatospora* isolates based upon six concatenated genes

Taking in to account the potential weakness of SSU-based phylogenies for discriminating closely related microsporidian (and other) taxa, phylogenies based upon alternative (coding) regions of the *Hepatospora* genome were constructed to investigate their potential as taxonomic discriminators (GenBank accession nos. of *Hepatospora* genes used: KU695715, KU695716, KU695717, KU695718, KU695719, KU695720, KU695721, KU695722, KU695723, KU695724, KU695725, KU695726, KU695727, KU695728, KU695729, KU695730, KU695731, KU695732). The resulting alignment of the six masked concatenated genes consisted of 18 232 sites. Phylogenetic trees based on ML and BI methods displayed identical topologies and strongly supported the grouping of the Chinese mitten crab parasite *H. eriocheir*, the edible crab and pea crab parasites, and the placement of *E. cuniculi* strains and *Nematocida* spp. as distinct clades with high confidence values (Fig. 4). The clade consisting of *H. eriocheir*, the edible crab parasite and the pea crab parasites branched as a sister group to *Enterocytozoon bieneusi* as consistent with previously phylogenies based upon SSU gene sequences (Stentiford *et al.* 2011). Our tree is also consistent with a previous multi-protein microsporidian phylogeny showing the Enterocytozoonidae forming a clade with *Vittaforma corneae* and with the *Nosema/Encephalitozoon* clade as a sister group to these (Nakjang *et al.* 2013). Within our tree, all resequenced strains are retrieved as clades with few

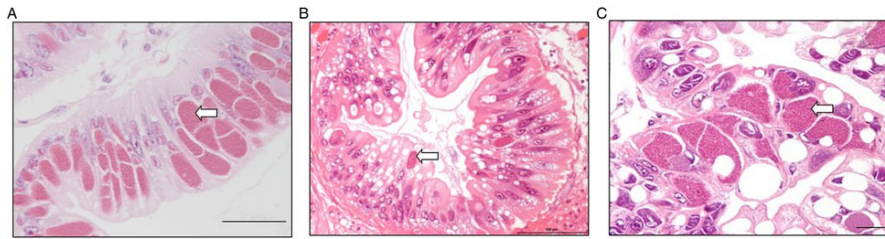


Fig. 1. (A) Histopathology of *Hepatospora* sp. in the hepatopancreas of *E. sinensis*. Scale = 50 μ m. (B) Histopathology of *H. eriocheir* in the hepatopancreas of *Cancer pagurus*. Scale = 100 μ m. (C) Histopathology of *Hepatospora* sp. in the hepatopancreas of *P. pisum*. Scale = 25 μ m. Tubule epithelial cells contained multiple granular inclusions (arrows) All images H&E histology.

nucleotide differences separating them. As observed for our crab pathogens, no nucleotide differences separated strains of *Nematocida* sp. or *Nematocida parisii*, in our analysis. Nucleotide differences do however separate *E. cucinuli* EC1-3 strains from *E. cucinuli* GB and *E. hellem* swiss from *E. hellem* ATCC. The six gene concatenated phylogeny supports the proposition from our SSU-based phylogeny that the parasites infecting Chinese mitten crabs, European edible crabs and, pea crabs can all be classified as *H. eriocheir*. Further, it provides firm evidence that *H. eriocheir* (or very closely related microvariants thereof) can infect a wide variety of decapods crustaceans from different aquatic habitats.

DISCUSSION

Phylogenetic analyses performed with rDNA have been pivotal in erecting new taxa and in the discovery of morphological plasticity within the microsporidian phylum but have however been unsuccessful in resolving the branching relationships between closely related species. The *Hepatospora* genus is a recent taxa erected as a result of modern phylogenetic techniques (Stentiford *et al.* 2011). It has been defined as a genus that encompasses microsporidia with infective life stages that develop within the hepatopancreas of marine and brackish water hosts. With the use of rDNA-based phylogenetic analyses, Stentiford *et al.* (2011) renamed and updated the description of *Endoreticulatus eriocheir* (Wang and Chen, 2007) to form the type species *H. eriocheir*. Unlike Wang and Chen who isolated their spores from native Chinese mitten crabs, the spores in the Stentiford *et al.* (2011) study were isolated from invasive Chinese mitten crabs (*E. sinensis*) which had been caught in the Thames estuary, UK. In both studies, the spores were described as ellipsoid, measuring $\sim 1.8 \times 0.9 \mu$ m and containing seven to eight polar filaments. Recently, Ding *et al.* (2016) have shown that not only is the pathogen infecting the

Asian population of Chinese mitten crabs *H. eriocheir*, but also the infection is associated with an emerging disease condition causing significant proportions in aquaculture production of this species in China. In their original taxonomic paper, Stentiford *et al.* (2011) also highlighted an unassigned microsporidian infecting the hepatopancreas of edible crabs, *C. pagurus*, and suggested at the time that this was also likely a member of the genus *Hepatospora* based upon high similarity with the partial SSU gene sequence from *H. eriocheir*.

In 2012, Longshaw *et al.* produced a disease profile of the pea crab (*P. pisum*) revealing the presence of two uncategorized microsporidian parasites. One of these appeared to be cytoplasmic, residing in hepatopancreatocytes and inducing a necrotizing effect that resulted in the degeneration of hepatopancreatic tubules. However, in this study, the microsporidian was not assigned a taxon and no further data was presented in order to aid its molecular or morphological characterization.

In the present study, we have shown that infection with the parasite of the hepatopancreas of edible crabs displays a similar pathology to that caused by infection with *H. eriocheir* in Chinese mitten crabs from London (Stentiford *et al.* 2011) and from China (Wang and Chen, 2007; Ding *et al.* 2016), and to the infection described in pea crabs by Longshaw *et al.* (2012). These parasites also appear to share a broadly similar morphological development within host gut epithelial cells albeit with some distinctive differences in karyo-status (*H. eriocheir* is unikaryotic whereas the pea crab and edible crab parasites are dikaryotic) and some minor differences in polar tube-coiling patterns (summarized in Table 1). Based upon our analysis of rDNA gene sequence-based phylogenetic relationship between *H. eriocheir* and the two novel parasites, all three are virtually indistinguishable, forming a monophyletic group immediately adjacent to the Enterocytozoonidae clade (Stentiford *et al.* 2013) (see Fig. 3).

Based upon the somewhat surprising finding that the same parasite taxon appeared to infect not only

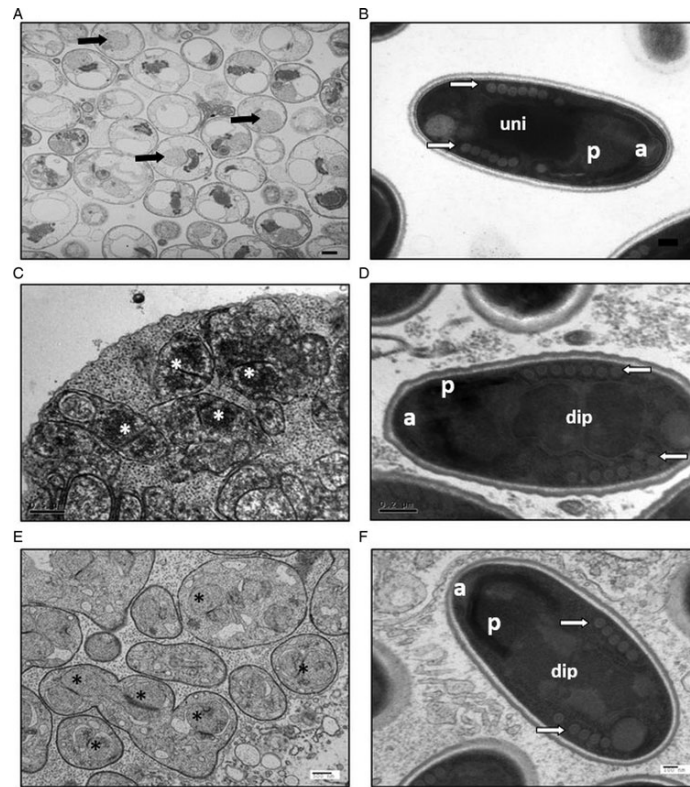


Fig. 2. Electron micrographs of *H. eriocheir* from *E. sinensis* (A and B) and *Hepatospora* sp. from *Cancer pagurus* (C and D) and *P. pisum* (E and F). (A) Developing sporonts within a parasitophorous vesicle. Sporonts contain multiple unikaryotic nuclei (arrows) which then divide and mature to form spores. Scale bar = 500 nm. (B) Mature spore containing seven to eight turns of the polar filament (arrow) in single file, unikaryotic nucleus (uni), anchoring disc (a) and polaroplast (p). Scale bar = 100 nm. (C) Developing sporonts within a parasitophorous vesicle. Sporonts contain multiple diplokaryotic nuclei (*) which then divide and mature to form spores. Scale bar = 0.5 μ m. (D) Mature spore containing five to six turns of the polar filament (arrow) in single file, diplokaryotic nucleus (dip), anchoring disc (a) and polaroplast (p). Scale bar = 0.2 μ m. (E) Developing sporonts within a parasitophorous vesicle. Sporonts contain multiple diplokaryotic nuclei (*) which then divide and mature to form spores. Scale bar = 500 nm. (F) Mature spore containing five to six turns of the polar filament (arrow) in single file, diplokaryotic nucleus (dip), anchoring disc (a) and polaroplast (p). Scale bar = 100 nm.

Chinese mitten crabs from different parts of their invasive/native range, but also occurred in niche-separated crab hosts from the European marine environment we used these three microsporidians as a model system to assess how a concatenated multi-gene phylogenetic approach could be applied as refined tool for discriminating taxa which cannot be separated by rDNA-based phylogenetic approaches. As outlined in our recent Opinion piece, appropriate application of such techniques may be suitable where significant biological reasoning exists for potential separation of taxa (e.g. one causing disease and another not) and, may be

particularly important where pathogens are to be listed to prevent their international trade in animals and their products (Stentiford *et al.* 2014). The model system presented by *Hepatospora* was appropriate given the potential for geographical, host and habitat distinction between known isolates, some reasonably significant differences in morphology (i.e. karyotstatus) and, the availability of draft genome data for the best studied of these, *H. eriocheir*. The rationale behind this concatenated phylogeny approach was to look for alternatives or additional data with which to resolve the branching relationship between very closely related

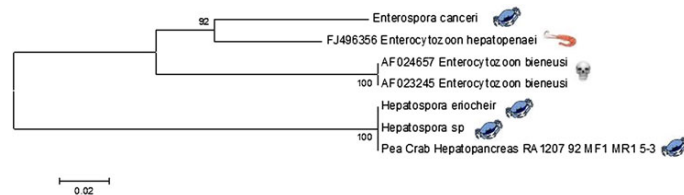


Fig. 3. Neighbour-joining tree based on a 511 bp nucleotide partial SSU 18S sequence from parasites isolated from hepatopancreas of *E. sinensis* (*H. eriocheir*), *Cancer pagurus* (*Hepatospora* sp.) and *P. pisum*. The phylogenetic analysis was performed using Mega version 5.05. Analysis was done using 1000 bootstrapped datasets and values >70% are shown on the tree. The scale bar represents substitutions per nucleotide site.

microsporidia. Despite this reasoning, our concatenated phylogenies resulted in almost identical tree topologies in both the ML and BI probabilistic approaches used with nodes supported by high bootstrap and posterior probability values, respectively. This, in addition to the retrieval of well-known relationships like the grouping of *Encephalitozoon* and *Nematocida* strains (in Fig. 4) and overall tree topology similar to previously published work based on rDNA genes (Troemel *et al.* 2008; Stentiford *et al.* 2011; Cuomo *et al.* 2012) increases our confidence in the phylogenetic relationships inferred by our study.

Whilst our multi-gene phylogeny confirms the close relationship of the three isolates, it does not discriminate well between them. This is in part due to the use of highly conserved genes for our concatenated tree phylogenetic analyses. Even though protein-coding genes have been successfully used in previous studies to infer deep phylogenetic relationships within the microsporidian phylum and placing microsporidians within the tree of life (Fast *et al.* 1999; Hirt *et al.* 1999; Keeling, 2003; Nakjang *et al.* 2013), the selected genes were unable to properly resolve branching relationships in our analyses due to the high level of nucleotide sequence similarity between the three *Hepatospora* strains (Table 1 and Fig. 3). Due to the small size of the pea crab hepatopancreas and very low infection levels in this host, we were unable to extract a large quantity of parasite genomic DNA. This in turn limited the number of marker genes we could amplify from the pea crab parasite for this study. Given more material we would have amplified further genes, for example polar tube and spore wall protein-coding genes. These have been used in previous studies to distinguish between *Encephalitozoon* and *Nosema* strains (Peuvel *et al.* 2000; Polonais *et al.* 2010; Chaimanee *et al.* 2011), and could have been added in our analyses to improve the resolution of branching relationships between the *Hepatospora* strains. It must however be noted that even though Polonais *et al.* (2010) were successful in differentiating between *Encephalitozoon hellem* strains by looking at nucleotide polymorphisms at the spore wall protein gene,

this gene was unsuccessful in differentiating between *Nosema ceranae* geographical isolates in similar recent studies (Roudel *et al.* 2013; Van der Zee *et al.* 2014). This highlights the different evolutionary pressures acting on the genes and possibly genomes of different microsporidians and that a gene successfully used to differentiate between strains in one microsporidian species may not be ideal for other species.

Future studies of this kind should probably focus on divergent single copy orthologues between strains of interest. Primer design for more variable genes is more challenging, but the advent of single-cell sequencing will remove this issue completely. As microsporidian genomes become increasingly available, we envisage a switch from phylogenetics to phylogenomics as the latter presents a more holistic approach to understanding close phylogenetic relationships and providing information for more robust taxonomic assignments.

Taxonomy of *Hepatospora*

Despite the caveats exposed in the preceding section, it is useful to consider the taxonomic placement of the parasites infecting the three host crabs. Here, it is useful to look to other genera within the phylum. Since already established subspecies of *E. cuniculi*, *E. hellem* and *Nematocida* (also included within our analysis) have nucleotide similarity for the same genes used in this study ranging between 92.13 and 99.95%, assigning different species names to the host-specific isolates of *Hepatospora* (98.53–98.88% nucleotide similarity) is not supported (see highlighted boxes in Table 4). On the bases of their high nucleotide similarity and their consequent grouping with minimal branching distances on our multi-gene phylogenetic tree (shown in bold in Fig. 4), we propose that *Hepatospora* parasites investigated in the current study should therefore be regarded as the same species, *H. eriocheir*, or potentially very closely related microvariants thereof. However, in consideration of the differences in morphological features that would have placed these microsporidians in completely different

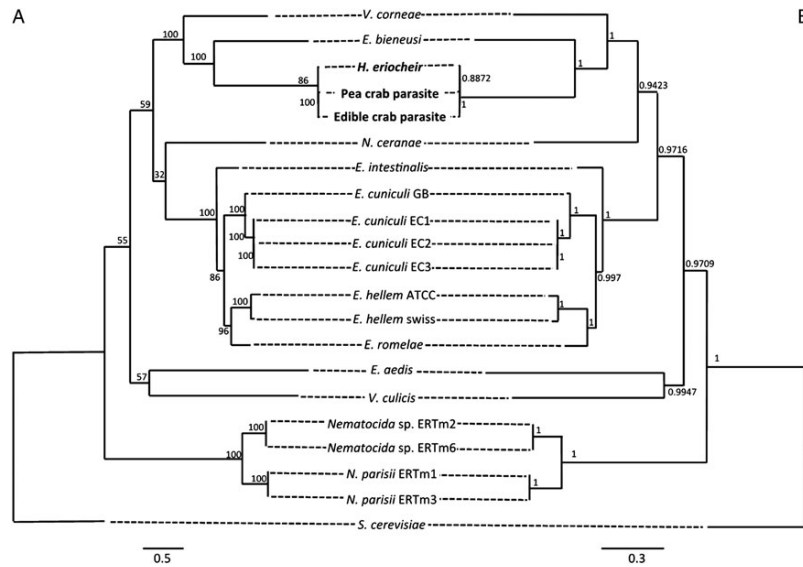


Fig. 4. Grouping of three *Hepatospora*/*Hepatospora*-like species suggests they are closely related. Phylogenetic trees based on (A.) maximum likelihood, (B.) Bayesian inference of 20 microsporidians for six concatenated genes rooted with *S. cerevisiae*. Numbers on nodes are (A.) Bootstrap confidence levels from 100 replicates, (B.) Bayesian posterior probability values. Both trees displaying identical topologies and grouping of *Hepatospora*/*Hepatospora*-like clade are shown in bold. The scale bars represent nucleotide substitutions per site.

Table 4. High nucleotide sequence similarity of the six marker genes used in this study between the parasites isolated from three different crab hosts

	arginyl tRNA synthetase	prolyl tRNA synthetase	beta-tubulin	chitin synthase	HSP70	RNA polymerase II
<i>H. eriocheir</i> vs Pea crab parasite	99 21/669	98 8/482	99 1/643	100 0/228	100 0/305	99 14/1341
<i>H. eriocheir</i> vs Edible crab parasite	99 23/1064	99 29/1401	99 6/1200	99 15/2211	99 6/744	99 22/3303
Edible crab parasite vs Pea crab parasite	99 20/669	99 1/482	99 1/643	100 0/228	99 4/305	99 15/1341

Numbers in bold are percentage identity comparison of sites (i.e. nucleotides + gaps) resulting from the alignment of each of the six marker genes of three *Hepatospora*/*Hepatospora*-like microsporidian species. Italicized numbers represent number of variable nucleotides in the pairwise alignment of two species without taking gaps into account (number given of variable nucleotides/total number of aligned nucleotides).

taxonomic ranks using traditional approaches (and considering that taxonomy of the phylum abides by rules laid down by the ICZN), it is perhaps appropriate to consider these microvariants as subspecies. We therefore propose the assignment of *H. eriocheir pinnotheres* and *H. eriocheir canceri* as subspecies of *H. eriocheir* infecting the pea crab and edible crab, respectively. Regardless of use of particular nomenclature, it is noteworthy that the genus *Hepatospora* (and perhaps specifically its type taxon *H. eriocheir*) may represent an example of a parasite cline, infecting the guts of one of the

most abundant host groups in our oceans, the crustaceans. The minor differences in rDNA-based and even concatenated phylogenies for *Hepatospora* may underlie a subtly shifting genome required for survival in hosts from different habitats. As more subspecies are discovered from variant hosts in different habitats, the concept of the parasite cline can be better studied.

An alternative hypothesis would reflect the potential that a previously host-specific parasite, *H. eriocheir* has been inadvertently introduced to European waters by its invasive host, the Chinese mitten crab; following

which it has subsequently switched to hosts which at least have niche overlap at some point in their life cycle (Ingle, 1980; Lawton, 1989; Clark *et al.* 1998; Becker and Türkay, 2010). The description of gut infecting microsporidian taxa from locations where such niche overlap are absent will address this key question.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Matt Longshaw and Mr Jamie Bojko for their assistance in obtaining samples of pea crabs.

FINANCIAL SUPPORT

This work was supported by the Department of Environment, Food and Rural Affairs (DEFRA) under Contract Number FB002 (to Dr Stephen Feist, Project Manager and liaison between Cefas and Defra).

REFERENCES

- Aurrecochea, C., Barreto, A., Brestelli, J., Brunk, B. P., Caler, E. V., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O. S., Heiges, M., Iodice, J., Kissinger, J. C., Kraemer, E. T., Li, W., Nayak, V., Pennington, C., Pinney, D. F., Pitts, B., Roos, D. S., Srinivasamoorthy, G., Stoeckert, C. J., Jr., Treatman, C. and Wang, H. (2011). AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. *Nucleic Acids Research* **39**, D612–D619.
- Ausubel, F. M., Brent, R., Kingston, R. B., Moore, D. D., Seidman, J. G., Smith, J. A. and Struhl, K. (2002). *Current Protocols in Molecular Biology*. John Wiley & Sons Inc.
- Bateman, K. S., Hicks, R. J. and Stentiford, G. D. (2011). Disease profiles differ between non-fished and fished populations of edible crab (*Cancer pagurus*) from a major commercial fishery. *ICES Journal of Marine Science* **68**, 2044–2052.
- Becker, C. and Türkay, M. (2010). Taxonomy and morphology of European pea crabs (Crustacea: Brachyura: Pinnotheridae). *Journal of Natural History* **44**, 1555–1575.
- Brown, J. R. and Doolittle, W. F. (1999). Gene descent, duplication, and horizontal transfer in the evolution of glutamyl- and glutamyl-tRNA synthetases. *Journal of Molecular Evolution* **49**, 485–495.
- Cali, A., Kotler, D. P. and Orenstein, J. M. (1993). *Septata intestinalis* N. G., N. Sp., an intestinal microsporidian associated with chronic diarrhea and dissemination in AIDS patients. *Journal of Eukaryotic Microbiology* **40**, 101–112.
- Canning, E. U. (1953). A new microsporidian, *Nosema locustae* n. sp., from the fat body of the African migratory locust, *Locusta migratoria migratoria* ides R. & F. *Parasitology* **43**, 287–290.
- Capella-Gutierrez, S., Silla-Martinez, J. M. and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973.
- Capella-gutiérrez, S., Marcet-houben, M. and Gabaldón, T. (2012). Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. *BMC Biology* **10**, 47.
- Chaimanee, V., Chen, Y., Pettis, J. S., Scott Cornman, R. and Chantawannakul, P. (2011). Phylogenetic analysis of *Nosema ceranae* isolated from European and Asian honeybees in Northern Thailand. *Journal of Invertebrate Pathology* **107**, 229–233.
- Clark, P. F., Rainbow, P. S., Robbins, R. S., Smith, B., Yeomans, W. E., Thomas, M. and Dobson, G. (1998). The alien Chinese mitten crab, *Eriocheir sinensis* (Crustacea: Decapoda: Brachyura), in the Thames catchment. *Journal of the Marine Biological Association* **78**, 1215–1221.
- Corradi, N. and Keeling, P. J. (2009). Microsporidia: a journey through radical taxonomic revisions. *Fungal Biology Reviews* **23**, 1–8.
- Cuomo, C. A., Desjardins, C. A., Bakowski, M. A., Goldberg, J., Ma, A. T., Becnel, J. J., Didier, E. S., Fan, L., Heiman, D. I., Levin, J. Z., Young, S., Zeng, Q. and Troemel, E. R. (2012). Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth. *Genome Research* **22**, 2478–2488.
- Darke, M. A. (1962). Absorption and transport of fluorescent brighteners by microorganisms. *Applied Microbiology* **10**, 387.
- Ding, Z.-F., Meng, Q.-G., Liu, H.-Y., Yuan, S., Zhang, F.-X., Sun, M.-L., Zhao, Y.-H., Shen, M.-F., Zhou, G., Pan, J.-L., Wang, W. and Xia, A.-J. (2016). First case of hepatopancreatic necrosis disease (HPND) in pond-reared Chinese mitten crab, *Eriocheir sinensis* associated with microsporidian. *Journal of Fish Diseases* doi: 10.1111/jfd.12437.
- Docker, M. F., Kent, M. L., Hervio, D. M. L., Khattra, J. S., Weiss, L. M., Cali, A. and Devlin, R. H. (1997). Ribosomal DNA sequence of *Nucleospora salmonis* Hedrick, Groff and Baxa, 1991 (Microsporea: Enterocytozooidae): Implications for phylogeny and nomenclature. *Journal of Eukaryotic Microbiology* **44**, 55–60.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797.
- Edlind, T., Visvesvara, G., Li, J. and Katiyar, S. (1994). Cryptosporidium and microsporidian beta-tubulin sequences: predictions of benzimidazole sensitivity and phylogeny. *Journal of Eukaryotic Microbiology* **41**, 385.
- Fast, N. M., Logsdon, J. M. and Doolittle, W. F. (1999). Phylogenetic analysis of the TATA box binding protein (TBP) gene from *Nosema locustae*: evidence for a microsporidia-fungi relationship and spliceosomal intron loss. *Molecular Biological Evolution* **16**, 1415–1419.
- Gouy, M., Guindon, S. and Gascuel, O. (2010). SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biological Evolution* **27**, 221–224.
- Gresoviak, S. J., Khattra, J. S., Nadler, S. A., Kent, M. L., Devlin, R. H., Vivares, C. P., Fuente, E. and Hedrick, R. P. (2000). Comparison of small subunit Ribosomal RNA gene and internal transcribed spacer sequences among isolates of the intranuclear microsporidian *Nucleospora salmonis*. *Journal of Eukaryotic Microbiology* **47**, 379–387.
- Hinkle, G., Morrison, H. G. and Sogin, M. L. (1997). Genes coding for reverse transcriptase, DNA-directed RNA polymerase, and chitin synthase from the microsporidian *Spraguea lophii*. *Biological Bulletin* **193**, 250–251.
- Hirt, R. P., Healy, B., Vossbrinck, C. R., Canning, E. U. and Embley, T. M. (1997). A mitochondrial Hsp70 orthologue in *Vairimorpha necatrix*: molecular evidence that microsporidia once contained mitochondria. *Current Biology* **7**, 995–998.
- Hirt, R. P., Logsdon, J. M., Healy, B., Dorey, M. W., Doolittle, W. F. and Embley, T. M. (1999). Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 580–585.
- Ingle, R. W. (1980). *British Crabs*. Oxford University Press and British Natural History Museum, London.
- Ironsides, J. E. (2013). Diversity and recombination of dispersed ribosomal DNA and protein coding genes in microsporidia. *PLoS ONE* **8**, e55878.
- James, T. Y., Pelin, A., Bonen, L., Ahrendt, S., Sain, D., Corradi, N. and Stajich, J. E. (2013). Shared signatures of parasitism and phylogenomics unite Cryptomycota and Microsporidia. *Current Biology* **23**, 1548–1553.
- Keeling, P. J. (2003). Congruent evidence from α -tubulin and β -tubulin gene phylogenies for a zygomycete origin of microsporidia. *Fungal Genetics and Biology* **38**, 298–309.
- Kent, M. L., Hervio, D. M. L., Docker, M. F. and Devlin, R. H. (1996). Taxonomy studies and diagnostic tests for Myxosporean and Microsporidian pathogens of Salmonid fishes utilising Ribosomal DNA sequence. *Journal of Eukaryotic Microbiology* **43**, 98S–99S.
- Lawton, P. (1989). Predatory interaction between the brachyuran crab *Cancer pagurus* and decapod crustacean prey. *Marine Ecology Progress Series* **52**, 169–179.
- Longshaw, M., Feist, S. W. and Bateman, K. S. (2012). Parasites and pathogens of the endosymbiotic pea crab (*Pinnotheres pisum*) from blue mussels (*Mytilus edulis*) in England. *Journal of Invertebrate Pathology* **109**, 235–242.
- Mathis, A., Weber, R. and Deplazes, P. (2005). Zoonotic potential of the microsporidia. *Clinical Microbiology Reviews* **18**, 423–445.
- Medlin, L., Elwood, H. J., Stickel, S. and Sogin, M. L. (1998). The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene* **71**, 91–499.
- Mount, D. W. (2007). Using the basic local alignment search tool (BLAST). 22 Cold Spring Harbour Protocols. Adapted from “Sequence Database Searching for Similar Sequences,” Chapter 6, in 2nd (eds David W. Mount), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA, 2004, doi: 10.1101/pdb.top17.
- Nakjang, S., Williams, T. A., Heinz, E., Watson, A. K., Foster, P. G., Sandra, K. M., Heaps, S. E., Hirt, R. P. and Martin Embley, T. (2013). Reduction and expansion in microsporidian genome evolution: new insights from comparative genomics. *Genome Biology Evolution* **5**, 2285–2303.

- O'Mahony, E. M., Tay, W. T. and Paxton, R. J. (2007). Multiple rRNA variants in a single spore of the microsporidian *Nosema bombi*. *Journal of Eukaryotic Microbiology* **54**, 103–109.
- Peuvrel, I., Delbac, F., Metenier, G., Peyret, P. and Vivares, C. P. (2000). Polymorphism of the gene encoding a major polar tube protein PTP1 in two microsporidia of the genus *Encephalitozoon*. *Parasitology* **121**(Pt 6), 581–587.
- Polonais, V., Mazet, M., Wawrzyniak, I., Texier, C., Blot, N., El Alaoui, H. and Delbac, F. (2010). The human microsporidian *Encephalitozoon hellem* synthesizes two spore wall polymorphic proteins useful for epidemiological studies. *Infection and Immunity* **78**, 2221–2230.
- Reynolds, E. S. (1963). The use of lead citrate at high pH as an electron-opaque stain in electron microscopy. *Journal of Cell Biology* **17**, 208–212.
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. and Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **3**, 539–542.
- Roudel, M., Aufaivre, J., Corbara, B., Delbac, F. and Blot, N. (2013). New insights on the genetic diversity of the honeybee parasite *Nosema ceranae* based on multilocus sequence analysis. *Parasitology* **140**, 1346–1356.
- Sak, B., Kvac, M., Petzelková, K., Kvetonová, D., Pomajbíková, K., Mulama, M., Kiyang, J. and Modrý, D. (2011). Diversity of microsporidia (Fungi: Microsporidia) among captive great apes in European zoos and African sanctuaries: evidence for zoonotic transmission? *Folia Parasitologica* **58**, 81–86.
- Shaddock, J. A., Meccoli, R. A., Davis, R. and Font, R. L. (1990). Isolation of a microsporidian from a human patient. *Journal of Infectious Diseases* **162**, 773–776.
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.
- Stentford, G. D., Bateman, K. S., Dubuffet, A., Chambers, E. and Stone, D. M. (2011). *Hepatospora eriocheir* (Wang and Chen, 2007) gen. et comb. nov. infecting invasive Chinese mitten crabs (*Eriocheir sinensis*) in Europe. *Journal of Invertebrate Pathology* **108**, 156–166.
- Stentford, G. D., Bateman, K. S., Feist, S. W., Chambers, E. and Stone, D. M. (2013). Plastic parasites: extreme dimorphism creates a taxonomic conundrum in the phylum Microsporidia. *International Journal of Parasitology* **43**, 339–352.
- Stentford, G. D., Feist, S. W., Stone, D. M., Peeler, E. J. and Bass, D. (2014). Policy, phylogeny, and the parasite. *Trends in Parasitology* **30**, 274–281.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2731–2739.
- Tay, W. T., O'Mahony, E. M. and Paxton, R. J. (2005). Complete rRNA gene sequences reveal that the microsporidium *Nosema bombi* infects diverse bumblebee (*Bombus* spp.) hosts and contains multiple polymorphic sites. *Journal of Eukaryotic Microbiology* **52**, 505–513.
- Tourtip, S., Wongtripop, S., Sritunyalucksana, K. S., Stentford, G. D., Bateman, K. S., Sriurairatana, S., Chayaburakul, K., Chavadej, K. and Withyachumnarnkul, B. (2009). *Enterocytozoon hepatopenaei* sp. nov. (Microspora: Enterocytozoonidae), a parasite of the black tiger shrimp *Penaeus monodon* (Decapoda: Penaeidae): fine structure and phylogenetic relationships. *Journal of Invertebrate Pathology* **102**, 21–29.
- Troemel, E. R., Félix, M.-A., Whiteman, N. K., Barrière, A. and Ausubel, F. M. (2008). Microsporidia are natural intracellular parasites of the nematode *Caenorhabditis elegans*. *PLoS Biology* **6**, e309.
- Van der Zee, R., Gómez-Moracho, T., Pisa, L., Sagastume, S., García-Palencia, P., Maside, X., Bartolomé, C., Martín-Hernández, R. and Higes, M. (2014). Virulence and polar tube protein genetic diversity of *Nosema ceranae* (Microsporidia) field isolates from Northern and Southern Europe in honeybees (*Apis mellifera iberiensis*). *Environmental Microbiology Reports* **6**, 401–413.
- Vávra, J. and Larsson, J. I. R. (2014). Structure of Microsporidia. In *Microsporidia: Pathogens of Opportunity* (ed. Weiss, L. M. and Beenel, J. J.), pp. 1–70. John Wiley & Sons, Inc., Chichester.
- Vavra, J. and Undeen, A. H. (1970). *Nosema algerae* n. sp. (Cnidospora, Microsporida) a Pathogen in a Laboratory Colony of *Anopheles stephensi* Liston (Diptera, Culicidae). *Journal of Protozoology* **17**, 240–249.
- Vossbrinck, C. R., Baker, M. D., Didier, E. S., Debrunner-Vossbrinck, B. A. and Shaddock, J. A. (1993). Ribosomal DNA sequences of *Encephalitozoon hellem* and *Encephalitozoon cuniculi*: species identification and phylogenetic construction. *Journal of Eukaryotic Microbiology* **40**, 354–362.
- Vossbrinck, C. R., Andreadis, T. G. and Debrunner-Vossbrinck, B. A. (1998). Verification of intermediate hosts in the life cycles of Microsporidia by small subunit rDNA sequencing. *Journal of Eukaryotic Microbiology* **45**, 290–292.
- Vossbrinck, C. R. and Debrunner-Vossbrinck, B. A. (2005). Molecular phylogeny of the Microsporidia: ecological, ultrastructural and taxonomic considerations. *Folia Parasitologica* **52**, 131–142.
- Wang, W. and Chen, J. (2007). Ultrastructural study on a novel microsporidian, *Endoreticulatus eriocheir* sp. nov. (Microsporida, *Encephalitozoonidae*), parasite of Chinese mitten crab, *Eriocheir sinensis* (Crustacea, Decapoda). *Journal of Invertebrate Pathology* **94**, 77–83.

Appendix 13: Wiredu-Boakye *et al.*, 2016 paper submitted to PLoS Pathogens

Manuscript

[Click here to download Manuscript EnterocytozoonidaeGenomicsBW.docx](#)

✎

1 **Comparative genomics of the *Enterocytozoonidae* reveals**
2 **decay of the glycolytic pathway and adaptations to intra-**
3 **nuclear parasitism**

4

5 Dominic Wiredu Boakye¹, Pattana Jaroenlak^{2,3}, Ay A. Prachumwat⁴, Tom A.
6 Williams⁵, Kelly S. Bateman⁶, Ornthuma Itsathitphisarn^{2,3}, Kallaya
7 Sritunyalucksana⁴, Konrad H. Paszkiewicz¹, Karen A. Moore¹, Grant D.
8 Stentiford^{6*}, Bryony A. P. Williams^{1*}

9

10 ¹ Biosciences, College of Life and Environmental Sciences, University of
11 Exeter, EX4 4QD, UK.

12 ² Department of Biochemistry, Faculty of Science, Mahidol University, Rama
13 VI Rd., Bangkok 10400, Thailand

14 ³ Center of Excellence for Shrimp Molecular Biology and Biotechnology,
15 Faculty of Science, Mahidol University, Rama VI Rd., Bangkok 10400,
16 Thailand

17 ⁴ Shrimp–Virus Interaction Laboratory (ASVI), National Center for Genetic
18 Engineering and Biotechnology (BIOTEC), Rama VI Rd., Bangkok 10400,
19 Thailand

20 ⁵ School of Earth Sciences, University of Bristol, BS8 1TH, UK

21 ⁶ European Union Reference Laboratory for Crustacean Diseases, Centre for
22 Environment Fisheries and Aquaculture Science, Weymouth Laboratory,
23 Weymouth, Dorset DT4 8UB, UK

24 *Correspondence should be addressed to b.a.p.williams@exeter.ac.uk or
25 grant.stentiford@cefasc.co.uk

1

1 **Abstract:**

2 The microsporidia are endoparasitic fungi living at the limits of cellular
3 streamlining, with genomes and metabolic capabilities reduced beyond those
4 of any other eukaryote. Microsporidia in the Enterocytozoonidae family are
5 unusual in terms of their hosts, host tissue preferences and levels of
6 metabolic reduction: All known taxa infect aquatic organisms apart from the
7 human pathogen, *Enterocytozoon bieneusi*; most known members live within
8 the host cytoplasm, but some live in close association with, or even within, the
9 host nucleus. Surprisingly, genome surveys of *E. bieneusi* revealed that this
10 species has even lost the glycolytic pathway. This is particularly remarkable
11 given that microsporidia cannot generate ATP via oxidative phosphorylation.

12 Here we present genome sequences from four novel members of the
13 Enterocytozoonidae that shed new light on the unusual biology of the group.
14 These include the first genome from an intranuclear eukaryotic pathogen, the
15 crab parasite *Enterospora canceri*, and the genome of *Enterocytozoon*
16 *hepatopenaei*, a major parasite of whiteleg shrimp that is currently spreading
17 rapidly across shrimp farms in Asia. Phylogenetic analysis confirms that the
18 human pathogen *E. bieneusi* originates from zoonotic transfer, as it is firmly
19 nested within a clade of microsporidians that infect aquatic crustaceans. We
20 survey the distribution of key metabolic pathways and transporter proteins
21 across the clade demonstrating that genes encoding glycolysis are lost
22 throughout the Enterocytozoonidae. We also show that the intranuclear
23 parasite *E. canceri* encodes roughly double the transporter repertoire of its
24 closest relatives, suggesting that expansion of existing transporter families
25 may underpin the adaptation to an intranuclear lifestyle.

1 Uniquely amongst eukaryotes, the Enterocytozoonidae have lost any known
2 means of producing ATP. Our findings raise major questions about how these
3 organisms fuel their parasitic lifecycle, from spore germination to invasion of
4 the host cell.

5 **Author Summary**

6 Microsporidia are ubiquitous pathogens of a range of different animals.
7 Whilst many infections are benign, in animals with impaired health, for
8 example immunocompromised humans and intensively farmed animals, they
9 may become problematic. One of the most important families of microsporidia
10 is the Enterocytozoonidae, which includes the common human infecting
11 species, *Enterocytozoon bieneusi* and the shrimp pathogen *Enterocytozoon*
12 *hepatopenaei*. The latter has emerged and spread rapidly across shrimp
13 farms in Asia and is a major threat to food security, slowing shrimp growth
14 and reducing protein yield. The Enterocytozoonidae are also interesting from
15 an evolutionary perspective. All microsporidia have lost the ability to generate
16 ATP energy in their mitochondria, the main site of energy generation in
17 complex cells. However, in 2010 it was reported that the microsporidian *E.*
18 *bieneusi* had also done away with glycolysis, leaving them with no means of
19 generating ATP. Here, we sequence the genomes of four more microsporidia
20 in the Enterocytozoonidae group. This provides a molecular resource for
21 studying this important group of pathogens, but also reveals that this whole
22 family of parasites has lost any intrinsic method of generating their own
23 energy, a situation unique amongst complex cells, rendering them utterly
24 reliant on host energy resources.

25

1 Introduction

2 The microsporidia are important emergent pathogens of almost all
3 known animal phyla in all major biomes [1]. The Enterocytozoonidae family is
4 home to two of the most economically important microsporidian species.
5 *Enterocytozoon bieneusi* is the most prevalent human infecting microsporidian
6 with some studies reported prevalences of 58% in immunocompetent
7 individuals. *Enterocytozoon hepatopenaei* is a recently emerged pathogen of
8 the farmed shrimp species *Penaeus monodon* and *Penaeus vannamei* and
9 severely retards the growth of the shrimp and has rapidly spread across
10 south-east Asia [2]. Microsporidia are also pre-eminent model systems for
11 understanding the processes of metabolic, cellular and genomic reduction in
12 eukaryotes, with genomes as small as 2.3 Mb and as few as 1,990 encoded
13 genes [3]. A large set of core eukaryotic genes were jettisoned early in the
14 evolutionary history of microsporidia [4]. This left them without a mitochondrial
15 genome and without the ability to generate ATP via oxidative phosphorylation
16 [4, 5], whilst their close relative *Mitosporidium daphniae* retains a
17 mitochondrial genome and components of the electron transport chain [6].
18 With the absence of oxidative phosphorylation in microsporidia, glucose is
19 only partially metabolised via glycolysis to release 7 % of its full ATP potential
20 [7]. However, this energy supply is complemented by the import of ATP
21 molecules from the host with the help of horizontally acquired ATP importers
22 [4, 8-10]. Thus, microsporidians are considered to be entirely reliant on
23 glycolysis and ATP import from their hosts for energy [6, 11]. One emerging
24 idea is that glycolysis occurs primarily in the spore stage but shuts down in
25 intracellular life stages, where the parasite can tap into the host's energy

1 pools [11]. This is based on data that demonstrate that the enzyme
2 responsible for the first ATP-forming reaction in glycolysis is enriched in the
3 cytosol of the extracellular spores in comparison to the cytosol of the
4 intracellular meront stages of *Trachipleistophora hominis* [11] and the
5 accumulation of AOX and glycerol-3-phosphate dehydrogenase, two enzymes
6 associated with the reoxidation of NADH for glycolysis, in the spores of
7 *Antonospora locustae* [12].

8
9 When the genome of the human infecting microsporidian *E. bienersi* was deeply
10 surveyed using whole-genome shotgun sequencing, there was a surprising
11 revelation that most genes involved in glycolysis were absent, in addition to
12 the loss of those involved in the oxidative phosphorylation pathway [13, 14].
13 This implies that this parasite relies entirely on ATP import from its host with
14 no obvious means of generating ATP, even in the extracellular spore stage
15 [13, 14].

16
17 As organisms with highly reduced metabolisms, microsporidians are highly
18 dependent on their hosts and have an expanded repertoire of transport
19 proteins that allow them to exploit the rich environment in the host cell
20 cytoplasm. Interestingly, meronts of *E. bienersi* form tight physical
21 associations with the host mitochondria, and this, in other microsporidian
22 species, has been shown to maximise the host ATP-harvesting efficiency of
23 the parasite [8, 15]. However, this association of microsporidian life stages
24 with host mitochondria is not universal across the group [8]. In fact some of
25 the Enterocytozoonidae, all infecting aquatic animal hosts (crustaceans, fish)

1 do not inhabit the host cytoplasm but specifically the host nucleus. These
2 species include *Enterospora canceri*, *Enterospora nucleophila*, several
3 members of the genus *Nucelospora*, and, *Desmoozon lepeophtherii*
4 (= *Paranucleospora theridion*) [16-23]. These species have no access to the
5 host mitochondria so cannot use this strategy to maximise ATP uptake.

6

7 Despite the economic and evolutionary importance of the
8 Enterocytozoonidae, at present, *E. bieneusi* is the only member of the group
9 with a sequenced genome. This leaves several intriguing unanswered
10 questions about this group. Firstly, what drove these taxa to inhabit the host
11 nucleus and, what are the metabolic benefits and parasite adaptations to life
12 in this unusual environment? Secondly, when was glycolysis lost in the
13 microsporidia and does *E. bieneusi* represent an anomaly within the phylum?

14

15 To better understand this intriguing parasite family we sequenced, annotated
16 and analysed the genomes of several members of the Enterocytozoonidae,
17 including the intranuclear crab parasite *Enterospora canceri*, the cytoplasmic
18 crab parasites *Hepatospora eriocheir* and *Hepatospora eriocheir canceri*, and,
19 the cytoplasmic shrimp parasite *Enterocytozoon hepatopenaei*. The latter is of
20 particular interest given its recent emergence and significant economic impact
21 on the global shrimp farming industry [1, 24]. This pathogen is reported to
22 have reached epidemic levels in shrimp farms across South East Asia, where
23 the infection results in slow shrimp growth [25]. For this reason, this pathogen
24 is currently a major threat to sustainable shrimp farming in Asia [25]. Thus,
25 our data provide a resource for understanding the metabolic strategy for

6

1 intranuclear existence, as well as a foundation for investigating host-parasite
2 interactions (and potential intervention strategies) for economically important
3 microsporidians such as *E. hepatopenaei*. Additionally, our genome data
4 reveals that the genes that encode glycolysis show a pattern of parallel decay
5 across the Enterocytozoonidae clade: All of these species have highly
6 degenerated glycolytic pathways, but different genes have been lost in each
7 case. Our results imply a common loss of the selective pressure to retain
8 glycolysis across these diverse species of microsporidia despite their
9 drastically different habitats, hosts and tissue preferences.

10

11 **Results and Discussion**

12 **Establishing the interrelationships of the Enterocytozoonidae using**
13 **multi protein phylogeny:** In order to investigate the relationships between
14 our studied species within the Enterocytozoonidae, we generated a 21-protein
15 concatenated phylogeny. This demonstrated the close relationships of the
16 crustacean parasites *Ent. canceri*, and *E. hepatopenaei* to the human
17 pathogen *E. bienewisi*, with *Hepatospora* forming an outgroup to these (Fig. 1
18 + 2 and S1 Fig.). This phylogeny implies that there has been a rapid change
19 in lifestyle within the group with drastic switches in both cellular localisation
20 (*E. hepatopenaei* and *E. bienewisi* are cytoplasmic, while *Ent. canceri*, is
21 nuclear) and, host type (*E. hepatopenaei* and *Ent. canceri* in aquatic
22 crustaceans, and *E. bienewisi* in humans, mammals and birds) over a
23 relatively short evolutionary timescale. The finding that *E. bienewisi* is
24 phylogenetically nested amongst microsporidians with aquatic hosts suggests

7

1 that this species has emerged to infect human and terrestrial animal hosts via
2 zoonotic transfer from an aquatic environment.

3

4 **Parallel decay of the glycolysis and other metabolic pathways across**

5 **the Enterocytozoonidae:** The discovery that *E. bieneusi* lacked a functional
6 glycolytic pathway opened up a number of questions about how these
7 organisms fuel their complex lifecycle [13, 14]. One of these was whether the
8 loss was unique to this species or had occurred more broadly within the
9 Microsporidia. Firstly we took a conserved protein set identified by Keeling et
10 al. [14]. We used OrthoMCL to cluster all publically available microsporidia
11 predicted open reading frames and clusters representing these conserved
12 gene sets were used to populate S2 Table. In a second approach, the same
13 conserved protein set was used to interrogate our predicted ORFs and our
14 raw contigs for the Enterocytozoonidae using BLASTP and TBLASTN. This
15 revealed that that all surveyed representatives of the Enterocytozoonidae
16 lacked multiple components of the glycolytic pathway. We find 4 glycolytic
17 enzymes retained in *Ent. canceri*, 2 in *E. hepatopenaei* and 4 in each of the *H.*
18 *eriocheir* genomes. Although none of these genomes were sequenced to
19 completion, it is unlikely that our observations can be explained by a failure to
20 sequence the missing components of glycolysis. Using the presence or
21 absence of a set of 44 broadly conserved genes (transcriptional control
22 proteins) to gauge genome completeness, we estimate that our assemblies
23 are 84-100% complete [*H. eriocheir canceri* (40/44), *H. eriocheir* (40/44), *Ent.*
24 *canceri* (42/44), *E. hepatopaenei* (43/44)] (S2 Table). Assuming a genome
25 completeness of between 84% and 100%, we calculate the probabilities of

8

1 observing at least this many glycolytic losses simply as the result of
2 incomplete sequencing (rather than true absences) to be 2.869×10^{-08} for *Ent.*
3 *canceri*, 6.975×10^{-10} for *E. hepatopenaei*, 8.629×10^{-05} for *H. eriocheir*
4 *canceri* and 8.629×10^{-05} for *H. eriocheir*. In addition, the genomes from our
5 two independently sequenced and assembled *Hepatospora eriocheir* sub-
6 species share the same pattern of glycolytic enzyme loss corroborating our
7 assumptions that these gene absences are true losses and not a
8 consequence of incomplete sequencing. Overall, the data suggest a common
9 absence of a functional glycolytic pathway in the Enterocytozoonidae. Our
10 data also show that this loss did not occur as a single ancestral event but
11 highlights a pattern of parallel decay of the pathway with different genes lost
12 in each lineage (Fig. 2). This suggests a relaxation of the selective pressure to
13 keep glycolysis at the base of the Enterocytozoonidae followed by pathway
14 decay via loss of different enzymes. What is common to each species is that
15 one or more genes encoding a hexokinase-like protein are retained, along
16 with genes coding for a handful of other components. However and crucially,
17 phosphoglycerate kinase and pyruvate kinase are consistently missing. These
18 catalyse the steps in glycolysis in which ATP is generated. This means that
19 these fragmented glycolytic pathways potentially represent ATP sinks that
20 generate ADP. This observation motivates the hypothesis that the remaining
21 glycolytic components might be responsible for generating a high intracellular
22 ADP/ATP ratio in these parasites. Microsporidia are dependent on ADP/ATP
23 translocases in their intracellular stages. In the Chlamydiae, another group of
24 intracellular parasites, it has been shown that nucleotide import via
25 ADP/ATP translocases is stimulated by a high ADP/ATP ratio which would

1 occur when the cells are in a low energy state [26]. Thus the high ADP/ATP
2 ratio generated by the broken glycolytic pathway might be responsible for
3 stimulating ATP import into the microsporidia.

4

5 Other core metabolic enzymes were identified and mapped onto the multi-
6 protein phylogeny. Whilst individual microsporidian lineages have lost
7 enzymes of different metabolic pathways independently, the
8 Enterocytozoonidae have undergone the most dramatic metabolic reduction
9 described to date. In addition to the loss of most glycolytic enzymes, they
10 have also lost the pentose phosphate pathway and most genes involved in
11 fatty acid and trehalose metabolism (Figs. 1 & 2).

12

13 **The nature of microsporidian hexokinases in non-glycolytic lineages:**

14 Within the microsporidian phylum hexokinase is found either in single or
15 multiple copies. Based on sequence similarity, we identified hexokinase-like
16 genes in the Enterocytozoonidae that were divergent in comparison to other
17 microsporidian hexokinases. This may reflect the fact that they may not be
18 functioning in the same pathway as those present in glycolytic microsporidia.
19 As seen previously in other microsporidia [4], our data shows a pattern of
20 independent gene duplications of these hexokinases across the
21 Enterocytozoonidae, including multiple genes in *E. bienersi*, single gene
22 copies in *Ent. canceri* and *E. hepatopenaei* and a duplication event at the
23 base of the *Hepatospora* lineage (Fig. 3). An attractive explanation for the
24 retention and expansion of the hexokinase gene family in these seemingly
25 non-glycolytic organisms is that the redundant copy is secreted to modify host

1 metabolism and increase metabolic rewards for the pathogen as has been
2 previously hypothesized to occur in other glycolytic microsporidia [27].
3 However, no signal sequences that would typically direct these proteins to the
4 secretory pathway were detected in any of the Enterocytozoonidae
5 hexokinase proteins using either SignalP or WOLFPSORT [28, 29].

6

7 Interestingly, a protein tyrosine phosphatase A (PTPA) domain was found to
8 be fused to the N-terminus of the single copy of the *Ent. canceri* divergent
9 hexokinase, the significance of which is unclear. Typically this domain would
10 have the capacity to remove a phosphate from a phosphorylated tyrosine. In
11 other organisms, proteins containing these domain are involved in the
12 regulation of the function of other proteins [30]. This may suggest that the
13 protein now performs a dual role as both a protein phosphatase and a
14 hexokinase, or that potentially the protein phosphatase regulates the function
15 of the fused hexokinase protein.

16

17 The divergent nature of the gene sequences of these hexokinases suggested
18 that they may no longer be functioning as typical hexokinases. To investigate
19 this, we compared amino acid sequence conservation across known active
20 sites between the new members of the Enterocytozoonidae sequenced in this
21 study, hexokinases from published microsporidian genomes and the
22 hexokinase 1 (HXK1) of *Saccharomyces cerevisiae*. In the cases where a
23 genome of a species encodes multiple hexokinase homologues, the additional
24 hexokinase copies had lost some of the active sites identified in *S. cerevisiae*
25 HXK1. We obtained the same results when considering residues that are not

1 at the active site, but that have been experimentally proven to be important for
2 the function of the *S. cerevisiae* HXK1 [31]. Again, in these additional
3 hexokinase copies, we observed that these key catalytic residues were
4 frequently substituted or even deleted (Fig. 3).

5

6 Taken together, these observations suggest that at least one hexokinase in
7 each member of the Enterocytozoonidae retains active sites indicating that it
8 is capable of performing a conserved hexokinase function. When multiple
9 copies are present, the additional copies may have diverged in function to
10 perform other activities. However in the absence of a functional glycolytic
11 pathway this role remains unknown.

12

13 **Inferring the role of the conserved hexokinase gene in the metabolically**
14 **reduced genomes of the Enterocytozoonidae:** Hexokinase is perhaps best
15 known for its role in glycolysis. However, this enzyme also plays a key role in
16 at least three independent pathways: dolichyl-P-mannose synthesis, and the
17 trehalose and pentose phosphate pathways (PPP). In glycolysis, hexokinase
18 catalyzes the first step, which phosphorylates glucose into glucose-6-
19 phosphate (G6P), a molecule not recognized by glucose transporters thereby
20 forcing it towards three possible fates [32]. The first involves further
21 breakdown in the glycolytic pathway to produce pyruvate with the concomitant
22 net release of 2 ATP molecules. In the second, G6P is channeled into the
23 trehalose metabolic pathway which culminates in the storage of two glucose
24 molecules as a single trehalose molecule. Finally, G6P can also be fed into
25 the pentose phosphate pathway (PPP), which results in the production of

12

1 ribose-5-phosphate, required for the synthesis of nucleic acids (Berg et al.
2 2006). Of those enzymes required for channeling G6P into any of those
3 pathways, only glucose-6-phosphate dehydrogenase (G6PD), the enzyme
4 responsible for ushering G6P into the PPP, is conserved in the
5 Enterocytozoonidae. However, although G6PD is present, downstream
6 enzymes required for functioning of the PPP or for trehalose metabolism are
7 scantily conserved across members of the Enterocytozoonidae which means
8 that it is unlikely that hexokinase feeds into either of these pathways (Fig. 4).

9
10 The retention of hexokinase and G6PD in the genomes of these parasites
11 despite the independent loss of other genes involved in glycolysis, PPP and
12 trehalose metabolism, suggests an alternative role for these two enzymes. In
13 this context, it is interesting to note that the reducing equivalent, NADPH,
14 released during the conversion of glucose-6-phosphate to 6-
15 phosphogluconolate by G6PD is key for lipid synthesis and scavenging
16 harmful reactive oxygen species (ROS). Considering that microsporidia often
17 develop in direct contact with the host cell cytoplasm, under constant
18 challenge from host-derived ROS, retention of enzymes that catalyze the
19 production of NADPH may contribute to protection of parasite life stages from
20 ROS damage. Retention of the ROS metabolic pathway in all analyzed
21 species in this study is consistent with this hypothesis (Fig. 4). The coupled
22 reactions of hexokinase and G6PD culminate in the production of 6-
23 phosphogluconolate, which more readily diffuses across membranes than
24 glucose-6-phosphate [33-36] Speculatively, it is possible that 6-
25 phosphogluconolate may diffuse across the plasma membrane of

1 microsporidia into the host cell cytoplasm. Since the dehydrogenation of G6P
2 to 6-phosphogluconolate catalyzed by G6PD is a rate-limiting step, 6-
3 phosphogluconolate that diffuses into the host cell will be forced to proceed
4 further down the PPP to produce ribose 5-phosphate. This ribose sugar is
5 utilized for the synthesis of ribonucleic acids. However, excess levels of ribose
6 5-phosphate are fed back into the glycolytic pathway to produce ATP, which
7 can be taken up by the microsporidian meront via specialized plasma
8 membrane transporters (see fig. 4). Another possible explanation for the fate
9 of G6P in the Enterocytozoonidae may be that specialized transporters
10 [Drug/Metabolite Transporter (DMT) family](see fig. 4) transport G6P across
11 the membrane into the host cell, trafficking it in to the host's PPP or glycolytic
12 pathway, and forcing their hosts to bypass a rate-limiting step of glycolysis.
13 The outcome of this would be an increased supply of ATP from host
14 mitochondria.

15

16 Our gene content analysis demonstrates that out of the four metabolic
17 pathways that hexokinase could potentially feed into, only the dolichyl-P-
18 mannose (DPM) synthesis pathway has a complete repertoire of enzymes
19 (See fig. 4) across representatives of the Enterocytozoonidae (See fig. 4).
20 Perhaps the most obvious explanation for the retention of the DPM pathway in
21 these metabolically reduced lineages could be due its essential role in the
22 production of building blocks for biosynthesis and protein glycosylation [37-
23 40]. These pathways require high ATP levels and concomitantly output high
24 levels of ADP. Therefore, in addition to providing building blocks for protein
25 glycosylation, the DPM pathway together with the nucleotide base synthesis

1 pathways (one of the three other pathways retained across all analyzed taxa)
2 could also be contributing to a high ADP/ATP ratio within the developing
3 microsporidian meront. As described above, we hypothesise that these high
4 ADP/ATP ratios may drive increased ATP import into the microsporidia via the
5 ATP/ADP translocases in the microsporidian plasma membrane.

6

7

8 **Absence of an intrinsic ATP generating mechanism in the**
9 **Enterocytozoonidae:** The extreme genome reduction observed in these
10 *Enterocytozoon* lineages suggest they are completely reliant on horizontally
11 acquired ATP/ADP translocases that are universal in the Microsporidia for
12 ATP acquisition from the host cell [6]. These would provide ATP for the
13 Enterocytozoonidae merogonial stages, but it does not explain how these
14 lineages acquire or generate ATP during their extracellular spore stage for the
15 presumably energetic process of spore germination [41].

16

17 To explore the possibility that microsporidia have retained alternative ATP
18 generation pathways, we used BLAST searches to look for enzymes involved
19 in alternative exergonic metabolic pathways such as mixed acid fermentation,
20 the Entner-Doudoroff pathway, and lipid and protein oxidation. However, none
21 of the microsporidian genomes surveyed had the hallmark enzymes of these
22 pathways either. In addition, each of these pathways feeds into the citric acid
23 cycle, which is absent in the Microsporidia. Therefore polar tube extrusion by
24 means of ATP produced via the above-mentioned metabolic pathways is
25 unlikely in our studied Enterocytozoonidae.

1

2 An alternative cell invasion strategy that has been proposed for other
3 microsporidia is internalization into the cell by phagocytosis rather than via the
4 polar tube [42]. There is also evidence to suggest that host cell types
5 parasitized by the presently studied Enterocytozoonidae lineages
6 (hepatopancreatic epithelial cells and mammalian enterocytes), have
7 phagocytic properties [43-46]. It is therefore possible that the spores of these
8 lineages could enter their respective host cells by phagocytosis. This would
9 bring the microsporidian into contact with the host ATP supply and if
10 ATP/ADP translocases were expressed at this time, they could provide the
11 spore with the ATP needed to germinate and to escape from the phagocytic
12 vesicle. As mentioned, above, a truncated glycolytic pathway could potentially
13 stimulate ATP uptake in the microsporidian by generating a high ADP/ATP
14 ratio. Parasitic spore phagocytosis by the host cell could also explain how *Ent.*
15 *canceri* enters the its subcellular niche, the nucleoplasm. That is, the non-
16 sporulated *Ent. canceri* spore is phagocytosed into the host cytoplasm and
17 from there it germinates directly in to the host nucleoplasm. However, an
18 alternative scenario where engulfed spores infect the host nucleus via an
19 unknown mechanism independent of polar tube extrusion is also plausible as
20 such an invasion strategy has been observed in members of the
21 Cryptomycota (Syn. Rozellomycota) [47-49]. Phagocytosis is typically
22 mediated by pathogen ligand binding to phagocytic receptor – if this is the
23 main means of entry into the host cell, then identifying and blocking the
24 appropriate ligand may block pathogen entry into the host cell and provide a
25 means of controlling infection.

16

1

2 **Comparing the plasma membrane transporter repertoire of nuclear and**
3 **cytoplasm-infecting microsporidians:** To investigate the adaptation of *Ent.*
4 *canceri* to its unique intranuclear niche, we compared the predicted
5 complement of membrane transporters across the Enterocytozoonidae to
6 determine whether the intranuclear *Ent. canceri* genome encoded a different
7 transporter repertoire to that found in sister genera inhabiting the cytoplasm.
8 Interestingly, the *Enterocytozoonidae* as a whole encode fewer homologues
9 of characterised membrane transport proteins than other sequenced
10 microsporidia (median 23, compared to 46 and 48 in the outgroups
11 *Encephalitozoon cuniculi* and *Trachipleistophora hominis*, respectively),
12 suggesting a reduction in transport repertoire in their common ancestor (S3
13 table). However, the branch leading to *Ent. canceri* experienced a substantial
14 gain in characterised membrane transporters, encoding a number (45
15 transporters) similar to that of other microsporidians. These lineage-specific
16 gains included additional representatives of the ABC, lipid (P-type ATPase),
17 choline (CTL), amino acid, nucleotide sugar (UAA) and cation transporter
18 families, raising the possibility that, against a background of extreme genome
19 reduction, this increased transporter repertoire is associated with adaptation
20 to the intranuclear environment.

21

22 **The predicted secretomes of the Enterocytozoonidae:** A key question
23 related to members of the Enterocytozoonidae is how these closely-related
24 organisms, with vastly different parasitic habits and host preferences, interact
25 with their hosts. To screen for proteins secreted by these parasites into their

17

1 host cells, we screened each of the predicted proteomes with SignalP 4.0 and
2 identified those proteins with appropriate signal sequences to allow for
3 translocation to the host cell cytoplasm (or nucleoplasm) [29]. We filtered out
4 those proteins with multiple transmembrane domains that may be anchored
5 into parasite membranes. Although this list is likely to include some proteins
6 more generally associated with the secretory pathway or cell wall
7 biosynthesis, we also expected to find potential effectors. Whilst dominated by
8 hypothetical proteins, the list also included proteins with conserved domains
9 allowing prediction of their function (S4 table). Predicted secreted proteins
10 present in more than one Enterocytozoonidae genus, and thus representing
11 candidate effectors include putative peptidyl-prolyl cis-trans isomerases and
12 glycosyltransferases, both protein families that have been identified as
13 effectors in other microbes [50, 51] Most importantly, secreted bacterial
14 glycosyltransferases have been demonstrated to inhibit host-cell apoptosis
15 [50]. The phagocytic route of infection proposed in this manuscript for these
16 microsporidian lineages imply an initial spore wall-host-cell binding. Since
17 such interactions are known to trigger host cell apoptotic responses [52],
18 microsporidian glycosyltransferase effectors responsible for blocking this
19 apoptotic machinery may be crucial in ensuring the preservation of the host-
20 cell and the completion of the parasite's life cycle.

21

22 Apart from *E. bienersi*, all other known taxa within the Enterocytozoonidae
23 are pathogens of aquatic animals [1]. Here, we sequenced the genomes of
24 four of these taxa, known to infect aquatic crustaceans. The study has
25 provided a major new resource for the investigation of host-pathogen

1 interactions, phylogeny, genetic reduction and pathogen evolution in arguably
2 the most economically and evolutionarily important families within the phylum
3 [1]. We provide the first genome sequence for the major yield-limiting shrimp
4 pathogen *E. hepatopenaei*. Its public availability will undoubtedly underpin the
5 development of new tools to diagnose, monitor and potentially mitigate the
6 negative impacts of this pathogen. The genome of *Ent. canceri* represents the
7 first from an obligately intranuclear eukaryotic pathogen and in particular,
8 offers an insight into genomic strategies for existence away from contact with
9 the host cell metabolic machinery normally exploited by most microsporidians.
10 As a whole our data provide resounding evidence of the consistent loss of
11 glycolysis into the Enterocytozoonidae. We demonstrate that glycolytic
12 enzymes were not lost in a single event in the ancestor of the group but rather
13 there has been a common loss of the selective pressure to retain glycolysis
14 followed by a parallel erosion of the pathway by differential loss of the
15 enzymes across the members of the genus. Microsporidia are already primary
16 models of metabolic reduction in eukaryotes The Enterocytozoonidae take
17 this reduction even further making them the only eukaryotic group to have
18 eliminated all canonical ATP-generating pathways. Whilst there are
19 precedents for loss of glycolysis amongst prokaryotes that form very close
20 host associations, for example *Mycoplasma hominis* and *Nanoarchaeum*
21 *equitans*, [53, 54], these organisms are not ever found in isolation from their
22 hosts. Loss of glycolysis makes the Enterocytozoonidae unique amongst
23 eukaryotes and leaves the enigma of how they complete their life cycle,
24 surviving out of association with the host and activating their spores to invade
25 host cells in the absence of their own energy generation system.

1

2 **Materials and Methods**

3 **Sampling of edible crabs for *Enterospora canceri* and *Hepatospora***

4 ***eriocheir canceri***: European edible crab adults (*Cancer pagurus*) were
5 purchased from local fishermen in Weymouth, UK (50°34'N, 2°22'W) in
6 January 2013. The hepatopancreases isolated from crabs infected with either
7 *Enterospora canceri* or *Hepatospora eriocheir* were crushed with a sterile
8 pestle and mortar in 1 x PBS. The homogenous mash was then filtered
9 through a 100 µm mesh followed by cell sieving through 40 µm filter and the
10 filtrate was further purified using a percoll density gradient.

11 **Purification of *Enterocytozoon hepatopenaei* spores: *E. hepatopenaei***

12 infected shrimp *Penaeus vannamei* were collected from commercial shrimp
13 ponds in Thailand in January 2015. The size of the shrimp varied from 10-15
14 grams. The hepatopancreases were dissected out, homogenised with a sterile
15 ground-glass pestle, and filtered through 100 µm and 40 µm cell strainers
16 respectively to remove tissue debris. The filtered suspension was further
17 purified by a percoll density gradient. *Purification of *Hepatopenaei eriocheir**
18 *spores*: The hepatopancreas of an infected Chinese mitten crab was passed
19 through a 200 µm mesh and the filtrate was then again passed through a 40
20 µm mesh. The filtrate was incubated in a 1 x PBS -0.05% Triton X-100
21 solution for 1 hr. The washed and pelleted filtrate was then further purified
22 using a percoll gradient.

23 **Genomic DNA extraction for sequencing: For *H. eriocheir canceri*, *E.***

24 *hepatopenaei* and *Ent. canceri*, aliquots of purified spores were ground for 10
25 minutes in liquid nitrogen. This step was repeated three times before

20

1 resuspending the resulting powder in 800 µl of phenol before proceeding with
2 a standard phenol chloroform extraction followed by an ethanol precipitation
3 as per Campbell et al 2014 [55]. For *H. eriocheir*, purified spores were
4 subjected to bead beating followed by phenol/chloroform extraction and
5 ethanol precipitation as per [55].

6 **Sequencing Protocols**

7 One *E. hepatopenaei* DNA sample was processed with NextFlex Rapid
8 protocol without PCR amplification (BIOO Scientific) and was sequenced
9 using the MiSeq v3 reagents 300 (PE). A second *E. hepatopenaei* DNA
10 sample was processed for PacBio Sequencing. was performed using PacBio
11 RS II technology of a 10kb library on a single SMRT cell by Macrogen, South
12 Korea. For *Hepatospora eriocheir* DNA a SPRIworks fragment library
13 (Beckman Coulter) was generated and this was sequenced on the HiSeq
14 2000 100 PE. *E. canceri* DNA was used to generate a SPRIworks fragment
15 library (Beckman Coulter), the first of these was sequenced using MiSeq v2
16 reagents 250 (PE) and the second using MiSeq v3 reagents 300 (PE).
17 *Hepatospora eriocheir canceri* DNA was used to generate a SPRIworks
18 fragment library (Beckman Coulter) which was sequenced using the MiSeq v2
19 reagents 250 (PE).

20 **Genome assembly and genome annotation: MiSeq and HiSeq data:**
21 FASTQC was used to identify and filter/trim reads with poor quality scores. A
22 total of 197,235,682 reads with a maximum length of 300 bp were used to
23 assemble the *E. hepatopenaei* genome with the Spades assembly pipeline
24 (v.3) (Bankevich et al. 2012). A total of 222,450,433 paired-end reads with a
25 maximum length of 99 bp were used to assemble the genome of *H. eriocheir*

1 with the Velvet package (v1.1) (Zerbino & Birney 2008). A total of 8,740,870
2 paired-end Illumina reads with a maximum length of 251 bp were used to
3 assemble the *H. eriocheir canceri* genome with the A5_miseq assembly
4 pipeline (Coil et al. 2015). A total of 4,632,816 reads with a maximum length
5 of 301 were used to assemble the *Ent. canceri* genome with the A5_miseq
6 assembly pipeline (Coil et al. 2015). All reads used had a mean quality score
7 above 32. Protein prediction for sequenced genomes was performed with the
8 MAKER annotation package [56]. To enable the program to make the best
9 possible predictions for the newly assembled genomes of *Ent. canceri*, *E.*
10 *hepatopenaei*, *H. eriocheir* and *H. eriocheir canceri*, the ab-initio protein
11 prediction program incorporated within MAKER was trained on the published
12 genome of *E. bieneusi* as per user guide.

13 **PacBio data:** Subreads were assembled *de novo* by Canu (version 1.3) [57]
14 with the default parameter setting (<http://canu.readthedocs.io/>). Open-reading
15 frames (ORFs), tRNAs and rRNAs were predicted using Prokka (version 1.11)
16 [58]. Contaminating bacterial sequences found in the initial assembled contigs
17 were removed using a combination of both BLASTP and BLASTX searches
18 for putative ORFs and by BLASTN searches for rRNAs against the NCBI nr
19 and nt databases, respectively. ORFs were annotated using both BLASTP
20 and BLASTX results.

21

22 **Identification of core metabolic genes**

23 In order to assess how the protein repertoire encoded by the genomes
24 of *Ent. canceri*, *E. hepatopenaei* and *Hepatospora* spp. sequenced in this
25 study compared to that of other microsporidians, open reading frames (ORFs)

1 of all four sequenced genomes and those from 19 publicly available
2 microsporidian species were grouped into orthologous families with the
3 command line program, ORTHOMCL [59] set on the following parameters:
4 MCL inflation=1.1 and Maximum weight =100. These protein clusters were
5 subsequently parsed for 381 core microsporidian proteins to assess
6 completeness of newly sequenced genomes. This protein set has been
7 previously used for a similar study by Keeling *et al.*, 2010 [14]. As a further
8 verification for the presence/absence of this conserved gene set, BLASTP
9 and TBLASTN searches were carried out on our newly predicted ORFs and
10 newly generated genomic contigs. This revealed further predicted proteins
11 that were not picked up by our ORTHOMCL strategy or that were not
12 predicted by our MAKER annotation.

13

14 **Identification of transporter proteins from five genomes**

15 ORFs from all four genomes sequenced in this study were screened
16 using the transmembrane domain prediction tool, TMHMM [60] to select
17 proteins with one or more transmembrane domains. Since the TMHMM
18 program is known to erroneously identify signal peptides as transmembrane
19 domains the protein set was also analysed with SIGNALP [29]. Proteins
20 predicted to have a single transmembrane domain that was also identified to
21 be a signal peptide were removed from the analysis. A BLASTP search with
22 the remaining protein set was carried out against the following databases:
23 SGD, TCDB and NCBI databases [61-63]. Where possible, a consensus
24 BLAST hit ID was selected for each putative transmembrane protein.
25 WOLFPSORT [28] was subsequently used to predict a subcellular localization

23

1 for these proteins. Predicted plasma membrane proteins that had four or more
2 transmembrane domains were selected for further analysis as potential
3 transport proteins.

4 ORFs from the newly sequenced genomes and 19 other microsporidian
5 species were clustered into orthologous groups with the ORTHMCL [59]
6 command line program set on the following parameters: MCL inflation=1.1
7 and Maximum weight =100. For each predicted plasma membrane
8 transporter, its orthologous cluster was manually parsed to recover orthologs
9 that may have been missed by the automated pipeline described above. The
10 majority of the final predicted plasma membrane transporters (227/235)
11 assigned to a functional group had four or more transmembrane domains. For
12 predicted plasma membrane transporters that were not assigned to any
13 functional group, only those with four or more transmembrane domains were
14 retained. Although 18 non-Enterocytozoonidae microsporidian species were
15 used in this analysis, only data for *Enc. cuniculi* and *T. hominis* were
16 displayed in the final results.

17

18 **Identification of secreted proteins/potential effectors**

19 ORFs from the newly sequenced genomes and those from 19 other
20 microsporidian species were parsed with the TMHMM command line tool [60]
21 to detect proteins with transmembrane domains. Proteins that were devoid of
22 transmembrane domains or only possessed a single transmembrane domain
23 were retained for further analyses. This protein set was passed to the signal
24 peptide prediction program, SIGNALP [29]. Signal peptide-containing proteins
25 were retained as potential secreted effectors. Finally, a Gene Ontology

24

1 assessment of the predicted secreted proteins was performed with the
2 graphical interphase version of BLAST2GO 3.2 [64].

3

4 **Phylogenetic analyses**

5 Homologs for 22 universally conserved proteins: Wrs1p, Taf10p, TFIIE,
6 Taf10p, Tfa2p, Sec62p, Abd1p, SPT16, Brn1p, Nob1p, Caf40p, Tfb2p, Bos1p,
7 Npl4p, Tma46p, Tfa1p, Clp1p, Spt5p, Sec63p, Pri2p and Enp2p were aligned
8 with MUSCLE (v3.8.31) [65] using the default settings and subsequently
9 masked with the automatic command line tool TrimAl (v1.2rev59) [66]. The
10 substitution model prediction option on MEGA (v6.06) [67] was used to
11 identify the best substitution models for the individual protein homolog sets.
12 These models were then used to perform Maximum likelihood analyses on the
13 individual homolog protein sets using the command line program, RaxML
14 (v7.2.7) [68]. These were pilot trees to check for unusually long-branch
15 lengths indicative of unlikely orthologs. The individual masked alignments of
16 the homolog protein sets were manually concatenated using Seaview (v4.5.4)
17 [69]. A GTR+GAMMA module on RaxML [70] was subsequently used to
18 construct the final multi-protein concatenated phylogenetic tree.

19

20 **Authors' contributions:** DWB, BAW, KB, GS, PJ and KM performed the
21 experiments. DWB, BAW, AAP, TAW and KP analyzed the data KS, and OI
22 collected material for the project. DWB, BAW, and GS drafted the manuscript.

23

24 **Acknowledgements:** We thank Prof. Tim Flegel for facilitating this
25 collaboration between Thailand and the UK and for assistance with spore

1 purification work done in Bangkok.

2

3 **Data availability:** The generated datasets have been deposited at NCBI
4 under the biosample numbers SAMN04550297, SAMN04508211 and
5 SAMN04589881

6

7 **Figures:**

8 Fig. 1: Metabolic profiling of microsporidian genomes. Filled squares
9 represent presence of a gene involved in a metabolic pathway whereas a
10 blank space represents absence of a gene. The phylogenetic positions in the
11 cladogram are derived from Maximum Likelihood analyses performed on a
12 concatenated alignment of 21 conserved proteins. Values represent levels of
13 bootstrap support (100 replicates) for the corresponding nodes.

14

15 Fig. 2: Independent loss of glycolytic enzymes and retention of the
16 hexokinase gene across the Enterocytozoonidae family. Filled circles
17 represent presence of a gene whereas empty circles represents absence of a
18 gene The loss of different glycolytic genes in different lineages suggests that
19 the selective pressure for the retention of glycolysis was lost at the base of
20 clade, leading to evolutionary decay of the pathway over time.

21

22 Fig. 3: Phylogeny of hexokinase reveals that it is duplicated in some
23 microsporidian lineages: The phylogenetic positions are derived from
24 Maximum Likelihood analyses performed on a masked alignment of
25 hexokinase proteins. Mammalian hexokinases were used to root the tree.

26

1 Bootstrap support values that are above 95% (100 replicates) are represented
 2 by black dots on the respective nodes. Similarity between *S. cerevisiae*'s
 3 hexokinase active sites and those of other lineages are represented with a
 4 heat map.

5

6 Fig. 4: Comparing the enzyme repertoire of key metabolic pathways between
 7 members of the Enterocytozoonidae family and other microsporidians (figure
 8 adapted from Nakjang et al. 2013 [4]) Homologs of *S. cerevisiae* metabolic
 9 genes (represented on grey background) from *E. bienersi*, *Enc cuniculi* and
 10 *T. hominis* were retrieved by doing a "transform by orthology" search for the
 11 *S. cerevisiae* genes on the EupathDB public database. Genes for *Enc.*
 12 *cuniculi* or *E. bienersi* were subsequently queried against the newly
 13 sequenced genomes using BlastP to identify their corresponding homologs.
 14 Thick black arrows represent those pathways found in all Enterocytozoonidae.
 15 Where there is differential loss of the pathway across the Enterocytozoonidae,
 16 presence of a gene in each species is represented by a different coloured
 17 arrow. Predicted number of plasma membrane transporter families and their
 18 respective substrates are also colour coded for each species of the
 19 Enterocytozoonidae.

20

21 **Tables:**

22

23 **Table 1: Genome statistics for the newly sequenced genomes.** * Gene set
 24 taken from [71]

Statistics	<i>Hepatospora</i>	<i>Hepatospora</i>	<i>Enterospora</i>	<i>Enterocytozoon</i>
------------	--------------------	--------------------	--------------------	-----------------------

27

	<i>eriocheir</i>	<i>eriocheir canceri</i>	<i>canceri</i>	<i>hepatopenaei</i>
Assembly size (Mb)	4.57	4.84	3.10	3.26
Contigs	1300	2344	537	64
Mean coverage (X)	4477.89	63.18	288	363
Contig N50 (bp)	17583	3349	11128	125008
GC content (%)	22.44	23.16	40.15	25.45
GC content coding (%)	25.58	25.41	41.95	27.81
Coding regions (%)	42.39	40.31	59.50	71.87
Splicing machinery *	7/29 genes	7/29 genes	7/29 genes	7/29 genes
Genes	2716	3058	2179	2540

1

2 **Supporting files:**

1 **S1 Figure:** Concatenated phylogeny using a conserved predicted protein set
2 showing branch lengths and bootstrap support levels (Predicted protein
3 sequences used - Wrs1p, Taf10p, TFIIIE, Taf10p, Tfa2p, Sec62p, Abd1p,
4 SPT16, Brn1p, Nob1p, Caf40p, Tfb2p, Bos1p, Npl4p, Tma46p, Tfa1p, Clp1p,
5 Spt5p, Sec63p, Pri2p and Enp2p).

6 **S2 Table:** Extended list of representative pathways in diverse microsporidian
7 genomes.

8 **S3 Table:** Comparing plasma membrane transporter complement between
9 members and non-members of the Enterocytozoonidae family.

10 **S4 Table:** IDs of predicted secreted protein for the Enterocytozoonidae

11

12 **References:**

13 1. Stentiford GD, Becnel JJ, Weiss LM, Keeling PJ, Didier ES, Williams
14 BA, et al. Microsporidia - Emergent Pathogens in the Global Food Chain.
15 Trends Parasitol. 2016;32(4):336-48. doi: 10.1016/j.pt.2015.12.004. PubMed
16 PMID: 26796229; PubMed Central PMCID: PMC4818719.

17 2. Newman SG. Microsporidian impacts shrimp production-industry efforts
18 address control, not eradication. Glob Aquac Advocate. 2015:16-7.

19 3. Corradi N, Pombert JF, Farinelli L, Didier ES, Keeling PJ. The
20 complete sequence of the smallest known nuclear genome from the
21 microsporidian *Encephalitozoon intestinalis*. Nature communications.
22 2010;1:77. Epub 2010/09/25. doi: 10.1038/ncomms1082. PubMed PMID:
23 20865802; PubMed Central PMCID: PMC4355639.

24 4. Nakjang S, Williams TA, Heinz E, Watson AK, Foster PG, Sendra KM,
25 et al. Reduction and expansion in microsporidian genome evolution: new

1 insights from comparative genomics. *Genome biology and evolution*.
2 2013;5(12):2285-303. doi: 10.1093/gbe/evt184. PubMed PMID: 24259309;
3 PubMed Central PMCID: PMC3879972.

4 5. Williams BA, Hirt RP, Lucocq JM, Embley TM. A mitochondrial remnant
5 in the microsporidian *Trachipleistophora hominis*. *Nature*.
6 2002;418(6900):865-9.

7 6. Haag KL, James TY, Pombert JF, Larsson R, Schaer TM, Refardt D, et
8 al. Evolution of a morphological novelty occurred before genome compaction
9 in a lineage of extreme parasites. *Proceedings of the National Academy of*
10 *Sciences of the United States of America*. 2014;111(43):15480-5. Epub
11 2014/10/15. doi: 10.1073/pnas.1410442111. PubMed PMID: 25313038;
12 PubMed Central PMCID: PMCPmc4217409.

13 7. Berg JM, Tymoczko JL, Stryer L, Stryer LB. *Biochemistry*. 6th ed. ed.
14 New York: W. H. Freeman ; Basingstoke : Palgrave [distributor]; 2007.

15 8. Hacker C, Howell M, Bhella D, Lucocq J. Strategies for maximizing
16 ATP supply in the microsporidian *Encephalitozoon cuniculi*: direct binding of
17 mitochondria to the parasitophorous vacuole and clustering of the
18 mitochondrial porin VDAC. *Cell Microbiol*. 2014;16(4):565-79. doi:
19 10.1111/cmi.12240. PubMed PMID: 24245785; PubMed Central PMCID:
20 PMC4233961.

21 9. Richards TA, Hirt RP, Williams BA, Embley TM. Horizontal gene
22 transfer and the evolution of parasitic protozoa. *Protist*. 2003;154(1):17-32.
23 Epub 2003/06/19. PubMed PMID: 12812367.

24 10. Tsoulos AD, Kunji ER, Goldberg AV, Lucocq JM, Hirt RP, Embley TM.
25 A novel route for ATP acquisition by the remnant mitochondria of

1 *Encephalitozoon cuniculi*. Nature. 2008;453(7194):553-6. Epub 2008/05/02.
2 doi: nature06903 [pii]
3 10.1038/nature06903. PubMed PMID: 18449191.

4 11. Heinz E, Williams TA, Nakjang S, Noël CJ, Swan DC, Goldberg AV, et
5 al. The Genome of the Obligate Intracellular Parasite *Trachipleistophora*
6 *hominis*: New Insights into Microsporidian Genome Dynamics and Reductive
7 Evolution. PLOS Pathog. 2012;8(10):e1002979. doi:
8 10.1371/journal.ppat.1002979.

9 12. Dolgikh VV, Senderskiy IV, Pavlova OA, Naumov AM, Beznoussenko
10 GV. Immunolocalization of an alternative respiratory chain in *Antonospora*
11 (*Paranosema*) *locustae* spores: mitosomes retain their role in microsporidian
12 energy metabolism. Eukaryot Cell. 2011;10(4):588-93. doi:
13 10.1128/EC.00283-10. PubMed PMID: 21296913; PubMed Central PMCID:
14 PMCPMC3127642.

15 13. Akiyoshi DE, Morrison HG, Lei S, Feng X, Zhang Q, Corradi N, et al.
16 Genomic survey of the non-cultivable opportunistic human pathogen,
17 *Enterocytozoon bieneusi*. PLoS Pathog. 2009;5(1):e1000261. Epub
18 2009/01/10. doi: 10.1371/journal.ppat.1000261. PubMed PMID: 19132089;
19 PubMed Central PMCID: PMC2607024.

20 14. Keeling PJ, Corradi N, Morrison HG, Haag KL, Ebert D, Weiss LM, et
21 al. The reduced genome of the parasitic microsporidian *Enterocytozoon*
22 *bieneusi* lacks genes for core carbon metabolism. Genome biology and
23 evolution. 2010;2:304-9. doi: 10.1093/gbe/evq022. PubMed PMID: 20624735;
24 PubMed Central PMCID: PMCPMC2942035.

- 1 15. Desportes I, Lecharpentier Y, Galian A, Bernard F, Cochandpriollet B,
2 Lavergne A, et al. Occurrence of a New Microsporidan: *Enterocytozoon*
3 *bieneusi* n. g., n. sp., in the Enterocytes of a Human Patient with AIDS.
4 Journal of Protozoology. 1985;32(2):250-4. PubMed PMID:
5 WOS:A1985AKJ3300004.
- 6 16. Chilmonczyk S, Cox WT, Hedrick RP. *Enterocytozoon salmonis* N. Sp.
7 - an Intranuclear Microsporidium from Salmonid Fish. Journal of Protozoology.
8 1991;38(3):264-9. doi: DOI 10.1111/j.1550-7408.1991.tb04440.x. PubMed
9 PMID: WOS:A1991FQ06400017.
- 10 17. Freeman MA, Kasper JM, Kristmundsson A. *Nucleospora cyclopteri* n.
11 sp., an intranuclear microsporidian infecting wild lumpfish, *Cyclopterus*
12 *lumpus* L., in Icelandic waters. Parasit Vectors. 2013;6:49. doi: 10.1186/1756-
13 3305-6-49. PubMed PMID: 23445616; PubMed Central PMCID:
14 PMCPMC3606367.
- 15 18. Freeman MA, Sommerville C. *Desmozoon lepeophtherii* n. gen., n. sp.,
16 (Microsporidia: Enterocytozoonidae) infecting the salmon louse
17 *Lepeophtheirus salmonis* (Copepoda: Caligidae). Parasite Vector. 2009;2. doi:
18 Artn 58
19 10.1186/1756-3305-2-58. PubMed PMID: WOS:000272705200001.
- 20 19. Hedrick RP, Groff JM, Baxa DV. *Enterocytozoon salmonis*
21 Chilmonczyk, Cox, Hedrick (Microsporea): an intranuclear microsporidium
22 from chinook salmon *Oncorhynchus tshawytscha*. Dis Aquat Organ.
23 1991;10(2):103-8. doi: 10.3354/dao10103. PubMed PMID:
24 WOS:A1991FG54500004.

- 1 20. Lom J, Dykova I. Ultrastructure of *Nucleospora secunda* n. sp
2 (Microsporidia), parasite of enterocytes of *Nothobranchius rubripinnis*. Eur J
3 Protistol. 2002;38(1):19-27. doi: Doi 10.1078/0932-4739-00844. PubMed
4 PMID: WOS:000175733000003.
- 5 21. Nylund S, Nylund A, Watanabe K, Arnesen CE, Karlsbakk E.
6 *Paranucleospora theridion* n. gen., n. sp (Microsporidia, Enterocytozoonidae)
7 with a Life Cycle in the Salmon Louse (*Lepeophtheirus salmonis*, Copepoda)
8 and Atlantic Salmon (*Salmo salar*). Journal of Eukaryotic Microbiology.
9 2010;57(2):95-114. doi: 10.1111/j.1550-7408.2009.00451.x. PubMed PMID:
10 WOS:000275098400001.
- 11 22. Stentiford GD, Bateman KS. *Enterospora* sp., an intranuclear
12 microsporidian infection of hermit crab *Eupagurus bernhardus*. Dis Aquat
13 Organ. 2007;75(1):73-8. doi: DOI 10.3354/dao075073. PubMed PMID:
14 WOS:000247163500008.
- 15 23. Stentiford GD, Bateman KS, Longshaw M, Feist SW. *Enterospora*
16 *canceri* n. gen., n. sp., intranuclear within the hepatopancreatocytes of the
17 European edible crab *Cancer pagurus*. Dis Aquat Organ. 2007;75(1):61-72.
18 doi: 10.3354/dao075061. PubMed PMID: WOS:000247163500007.
- 19 24. Rajendran KV, Shivam S, Praveena PE, Rajan JJS, Kumar TS, Avunje
20 S, et al. Emergence of *Enterocytozoon hepatopenaei* (EHP) in farmed
21 *Penaeus (Litopenaeus) vannamei* in India. Aquaculture. 2016;454:272-80.
22 doi: 10.1016/j.aquaculture.2015.12.034. PubMed PMID:
23 WOS:000368690000034.
- 24 25. Thitamadee S, Prachumwat A, Srisala J, Jaroenlak P, Salachan PV,
25 Sritunyalucksana K, et al. Review of current disease threats for cultivated

1 penaeid shrimp in Asia. *Aquaculture*. 2016;452:69-87. doi:
2 <http://dx.doi.org/10.1016/j.aquaculture.2015.10.028>.

3 26. Trentmann O, Horn M, van Scheltinga AC, Neuhaus HE, Haferkamp I.
4 Enlightening energy parasitism by analysis of an ATP/ADP transporter from
5 chlamydiae. *PLoS Biol*. 2007;5(9):e231. doi: 10.1371/journal.pbio.0050231.
6 PubMed PMID: 17760504; PubMed Central PMCID: PMCPMC1951785.

7 27. Cuomo CA, Desjardins CA, Bakowski MA, Goldberg J, Ma AT, Becnel
8 JJ, et al. Microsporidian genome analysis reveals evolutionary strategies for
9 obligate intracellular growth. *Genome Res*. 2012;22(12):2478-88. doi:
10 10.1101/gr.142802.112. PubMed PMID: 22813931; PubMed Central PMCID:
11 PMCPMC3514677.

12 28. Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ,
13 et al. WoLF PSORT: protein localization predictor. *Nucleic Acids Research*.
14 2007;35:W585-W7. doi: 10.1093/nar/gkm259. PubMed PMID:
15 WOS:000255311500109.

16 29. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0:
17 discriminating signal peptides from transmembrane regions. *Nature Methods*.
18 2011;8(10):785-6. doi: 10.1038/nmeth.1701. PubMed PMID:
19 WOS:000295358000004.

20 30. Tonks NK. Protein tyrosine phosphatases: from genes, to function, to
21 disease. *Nat Rev Mol Cell Biol*. 2006;7(11):833-46. doi: 10.1038/nrm2039.
22 PubMed PMID: 17057753.

23 31. Kuser P, Cupri F, Bleicher L, Polikarpov I. Crystal structure of yeast
24 hexokinase PI in complex with glucose: A classical "induced fit" example

1 revised. Proteins. 2008;72(2):731-40. doi: 10.1002/prot.21956. PubMed
2 PMID: 18260108.

3 32. Mueckler M. Facilitative glucose transporters. Eur J Biochem.
4 1994;219(3):713-25. PubMed PMID: 8112322.

5 33. Becker JU, Betz A. Membrane transport as controlling pacemaker of
6 glycolysis in *Saccharomyces carlsbergensis*. Biochim Biophys Acta.
7 1972;274(2):584-97. PubMed PMID: 4340264.

8 34. Poincelot RP. Transport of Metabolites across Isolated Envelope
9 Membranes of Spinach Chloroplasts. Plant Physiol. 1975;55(5):849-52.
10 PubMed PMID: 16659179; PubMed Central PMCID: PMC541721.

11 35. Clifton D, Walsh RB, Fraenkel DG. Functional-Studies of Yeast
12 Glucokinase. J Bacteriol. 1993;175(11):3289-94. PubMed PMID:
13 WOS:A1993LE43100006.

14 36. Smits HP, Smits GJ, Postma PW, Walsh MC, VanDam K. High-affinity
15 glucose uptake in *Saccharomyces cerevisiae* is not dependent on the
16 presence of glucose-phosphorylating enzymes. Yeast. 1996;12(5):439-47.
17 doi: Doi 10.1002/(Sici)1097-0061(199604)12:5<439::Aid-Yea925>3.0.Co;2-W.
18 PubMed PMID: WOS:A1996UJ30000003.

19 37. Shental-Bechor D, Levy Y. Effect of glycosylation on protein folding: a
20 close look at thermodynamic stabilization. Proceedings of the National
21 Academy of Sciences of the United States of America. 2008;105(24):8256-61.
22 doi: 10.1073/pnas.0801340105. PubMed PMID: 18550810; PubMed Central
23 PMCID: PMC2448824.

24 38. Moharir A, Peck SH, Budden T, Lee SY. The role of N-glycosylation in
25 folding, trafficking, and functionality of lysosomal protein CLN5. PLoS One.

1 2013;8(9):e74299. doi: 10.1371/journal.pone.0074299. PubMed PMID:
2 24058541; PubMed Central PMCID: PMCPMC3769244.

3 39. Vagin O, Kraut JA, Sachs G. Role of N-glycosylation in trafficking of
4 apical membrane proteins in epithelia. *Am J Physiol Renal Physiol*.
5 2009;296(3):F459-69. doi: 10.1152/ajprenal.90340.2008. PubMed PMID:
6 18971212; PubMed Central PMCID: PMCPMC2660186.

7 40. Wujek P, Kida E, Walus M, Wisniewski KE, Golabek AA. N-
8 glycosylation is crucial for folding, trafficking, and stability of human tripeptidyl-
9 peptidase I. *J Biol Chem*. 2004;279(13):12827-39. doi:
10 10.1074/jbc.M313173200. PubMed PMID: 14702339.

11 41. Vavra J, Lukes J. Microsporidia and 'the art of living together'. *Adv*
12 *Parasitol*. 2013;82:253-319. doi: 10.1016/B978-0-12-407706-5.00004-6.
13 PubMed PMID: 23548087.

14 42. Franzen C, Muller A, Hartmann P, Salzberger B. Cell invasion and
15 intracellular fate of *Encephalitozoon cuniculi* (Microsporidia). *Parasitology*.
16 2005;130(Pt 3):285-92. PubMed PMID: 15796011.

17 43. Alday-Sanz V, Roque A, Turnbull JF. Clearing mechanisms of *Vibrio*
18 *vulnificus* biotype I in the black tiger shrimp *Penaeus monodon*. *Dis Aquat*
19 *Organ*. 2002;48(2):91-9. doi: 10.3354/dao048091. PubMed PMID: 12005240.

20 44. Bu HF, Wang X, Tang Y, Koti V, Tan XD. Toll-like receptor 2-mediated
21 peptidoglycan uptake by immature intestinal epithelial cells from apical side
22 and exosome-associated transcellular transcytosis. *J Cell Physiol*.
23 2010;222(3):658-68. doi: 10.1002/jcp.21985. PubMed PMID: 20020500;
24 PubMed Central PMCID: PMCPMC4414048.

- 1 45. Liu D, Yang J, Wang L. Cadmium induces ultrastructural changes in
2 the hepatopancreas of the freshwater crab *Sinopotamon henanense*. *Micron*.
3 2013;47:24-32. doi: 10.1016/j.micron.2013.01.002. PubMed PMID: 23402952.
- 4 46. Neal MD, Leaphart C, Levy R, Prince J, Billiar TR, Watkins S, et al.
5 Enterocyte TLR4 mediates phagocytosis and translocation of bacteria across
6 the intestinal barrier. *J Immunol*. 2006;176(5):3070-9. PubMed PMID:
7 16493066.
- 8 47. Corsaro D, Walochnik J, Venditti D, Muller KD, Hauröder B, Michel R.
9 Rediscovery of *Nucleophaga amoebae*, a novel member of the
10 Rozellomycota. *Parasitol Res*. 2014;113(12):4491-8. doi: 10.1007/s00436-
11 014-4138-8. PubMed PMID: 25258042.
- 12 48. Michel R, Schmid EN, Boker T, Hager DG, Muller KD, Hoffmann R, et
13 al. *Vannella* sp. harboring Microsporidia-like organisms isolated from the
14 contact lens and inflamed eye of a female keratitis patient. *Parasitol Res*.
15 2000;86(6):514-20. PubMed PMID: 10894481.
- 16 49. Michel R, Müller K-D, Hauröder B. A novel microsporidian endoparasite
17 replicating within the nucleus of *Saccamoeba limax* isolated from a pond.
18 *Endocytobios Cell Res*. 2009;19:120–6.
- 19 50. Lu Q, Li S, Shao F. Sweet Talk: Protein Glycosylation in Bacterial
20 Interaction With the Host. *Trends Microbiol*. 2015;23(10):630-41. doi:
21 10.1016/j.tim.2015.07.003. PubMed PMID: 26433695.
- 22 51. Unal CM, Steinert M. Microbial peptidyl-prolyl cis/trans isomerases
23 (PPlases): virulence factors and potential alternative drug targets. *Microbiol*
24 *Mol Biol Rev*. 2014;78(3):544-71. doi: 10.1128/MMBR.00015-14. PubMed
25 PMID: 25184565; PubMed Central PMCID: PMC4187684.

- 1 52. Pearson JS, Giogha C, Ong SY, Kennedy CL, Kelly M, Robinson KS,
2 et al. A type III effector antagonizes death receptor signalling during bacterial
3 gut infection. *Nature*. 2013;501(7466):247-51. doi: 10.1038/nature12524.
4 PubMed PMID: 24025841; PubMed Central PMCID: PMC3836246.
- 5 53. Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, Barnstead M, et
6 al. The genome of *Nanoarchaeum equitans*: insights into early archaeal
7 evolution and derived parasitism. *Proceedings of the National Academy of*
8 *Sciences of the United States of America*. 2003;100(22):12984-8. doi:
9 10.1073/pnas.1735403100. PubMed PMID: 14566062; PubMed Central
10 PMCID: PMC240731.
- 11 54. Pereyre S, Sirand-Pugnet P, Beven L, Charron A, Renaudin H, Barre
12 A, et al. Life on arginine for *Mycoplasma hominis*: clues from its minimal
13 genome and comparison with other human urogenital mycoplasmas. *PLOS*
14 *Genet*. 2009;5(10):e1000677. doi: 10.1371/journal.pgen.1000677. PubMed
15 PMID: 19816563; PubMed Central PMCID: PMC2751442.
- 16 55. Campbell SE, Williams TA, Yousuf A, Soanes DM, Paszkiewicz KH,
17 Williams BA. The genome of *Spraguea lophii* and the basis of host-
18 microsporidian interactions. *PLoS Genet*. 2013;9(8):e1003676. doi:
19 10.1371/journal.pgen.1003676. PubMed PMID: 23990793; PubMed Central
20 PMCID: PMC3749934.
- 21 56. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al.
22 MAKER: An easy-to-use annotation pipeline designed for emerging model
23 organism genomes. *Genome Research*. 2008;18(1):188-96. doi:
24 10.1101/gr.6743907. PubMed PMID: WOS:000251965300019.

- 1 57. Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM.
2 Assembling large genomes with single-molecule sequencing and locality-
3 sensitive hashing. *Nat Biotechnol.* 2015;33(6):623-30. doi: 10.1038/nbt.3238.
4 PubMed PMID: 26006009.
- 5 58. Seemann T. Prokka: rapid prokaryotic genome annotation.
6 *Bioinformatics.* 2014;30(14):2068-9. doi: 10.1093/bioinformatics/btu153.
7 PubMed PMID: 24642063.
- 8 59. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of ortholog
9 groups for eukaryotic genomes. *Genome Research.* 2003;13(9):2178-89. doi:
10 10.1101/gr.1224503. PubMed PMID: WOS:000185085300021.
- 11 60. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting
12 transmembrane protein topology with a hidden Markov model: Application to
13 complete genomes. *Journal of Molecular Biology.* 2001;305(3):567-80. doi:
14 10.1006/jmbi.2000.4315. PubMed PMID: WOS:000167760800017.
- 15 61. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan
16 ET, et al. *Saccharomyces* Genome Database: the genomics resource of
17 budding yeast. *Nucleic Acids Research.* 2012;40(D1):D700-D5. doi:
18 10.1093/nar/gkr1029. PubMed PMID: WOS:000298601300106.
- 19 62. Saier MH, Jr., Tran CV, Barabote RD. TCDB: the Transporter
20 Classification Database for membrane transport protein analyses and
21 information. *Nucleic Acids Research.* 2006;34:D181-D6. doi:
22 10.1093/nar/gkj001. PubMed PMID: WOS:000239307700038.
- 23 63. Tatusova T, Ciufu S, Fedorov B, O'Neill K, Tolstoy I. RefSeq microbial
24 genomes database: new representation and annotation strategy. *Nucleic*

1 Acids Research. 2014;42(D1):D553-D9. doi: 10.1093/nar/gkt1274. PubMed
2 PMID: WOS:000331139800082.

3 64. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M.
4 Blast2GO: a universal tool for annotation, visualization and analysis in
5 functional genomics research. *Bioinformatics*. 2005;21(18):3674-6. doi:
6 10.1093/bioinformatics/bti610. PubMed PMID: 16081474.

7 65. Edgar RC. MUSCLE: a multiple sequence alignment method with
8 reduced time and space complexity. *BMC Bioinformatics*. 2004;5:1-19. doi:
9 10.1186/1471-2105-5-113. PubMed PMID: WOS:000223920500001.

10 66. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for
11 automated alignment trimming in large-scale phylogenetic analyses.
12 *Bioinformatics*. 2009;25(15):1972-3. doi: 10.1093/bioinformatics/btp348.
13 PubMed PMID: WOS:000268107100022.

14 67. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S.
15 MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum
16 Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.
17 *Molecular Biology and Evolution*. 2011;28(10):2731-9. doi:
18 10.1093/molbev/msr121. PubMed PMID: WOS:000295184200003.

19 68. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and
20 post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312-3. doi:
21 10.1093/bioinformatics/btu033. PubMed PMID: WOS:000336095100024.

22 69. Gouy M, Guindon S, Gascuel O. SeaView Version 4: A Multiplatform
23 Graphical User Interface for Sequence Alignment and Phylogenetic Tree
24 Building. *Molecular Biology and Evolution*. 2010;27(2):221-4. doi:
25 10.1093/molbev/msp259. PubMed PMID: WOS:000273704400003.

- 1 70. Stamatakis A. RAxML-VI-HPC: Maximum likelihood-based
2 phylogenetic analyses with thousands of taxa and mixed models.
3 Bioinformatics. 2006;22(21):2688-90. doi: 10.1093/bioinformatics/btl446.
4 PubMed PMID: WOS:000241629600016.
- 5 71. Desjardins CA, Sanscrainte ND, Goldberg JM, Heiman D, Young S,
6 Zeng Q, et al. Contrasting host-pathogen interactions and genome evolution
7 in two generalist and specialist microsporidian pathogens of mosquitoes.
8 Nature communications. 2015;6:7121. doi: 10.1038/ncomms8121. PubMed
9 PMID: 25968466; PubMed Central PMCID: PMC4435813.
- 10

Figure 1

[Click here to download Figure Figure1.tif](#)

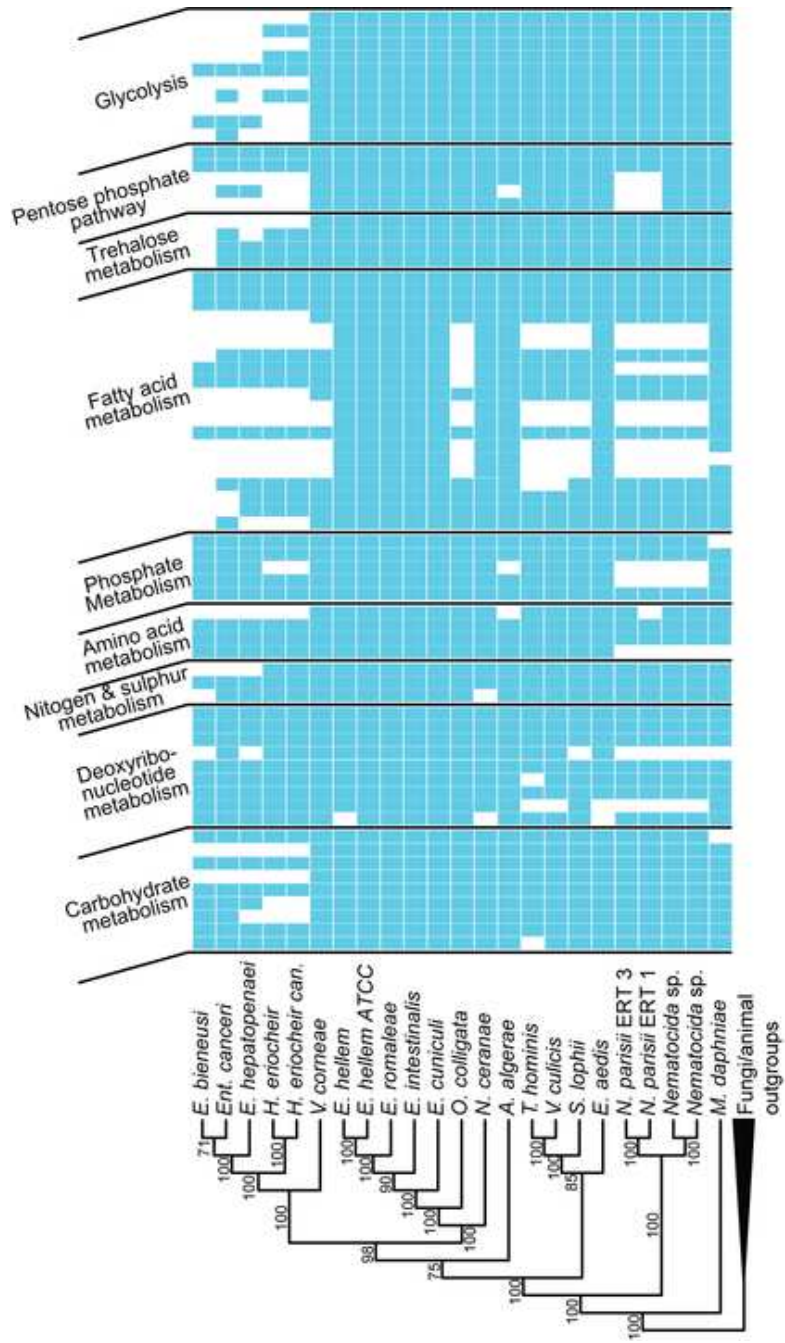


Figure 2

[Click here to download Figure Figure2.tif](#)

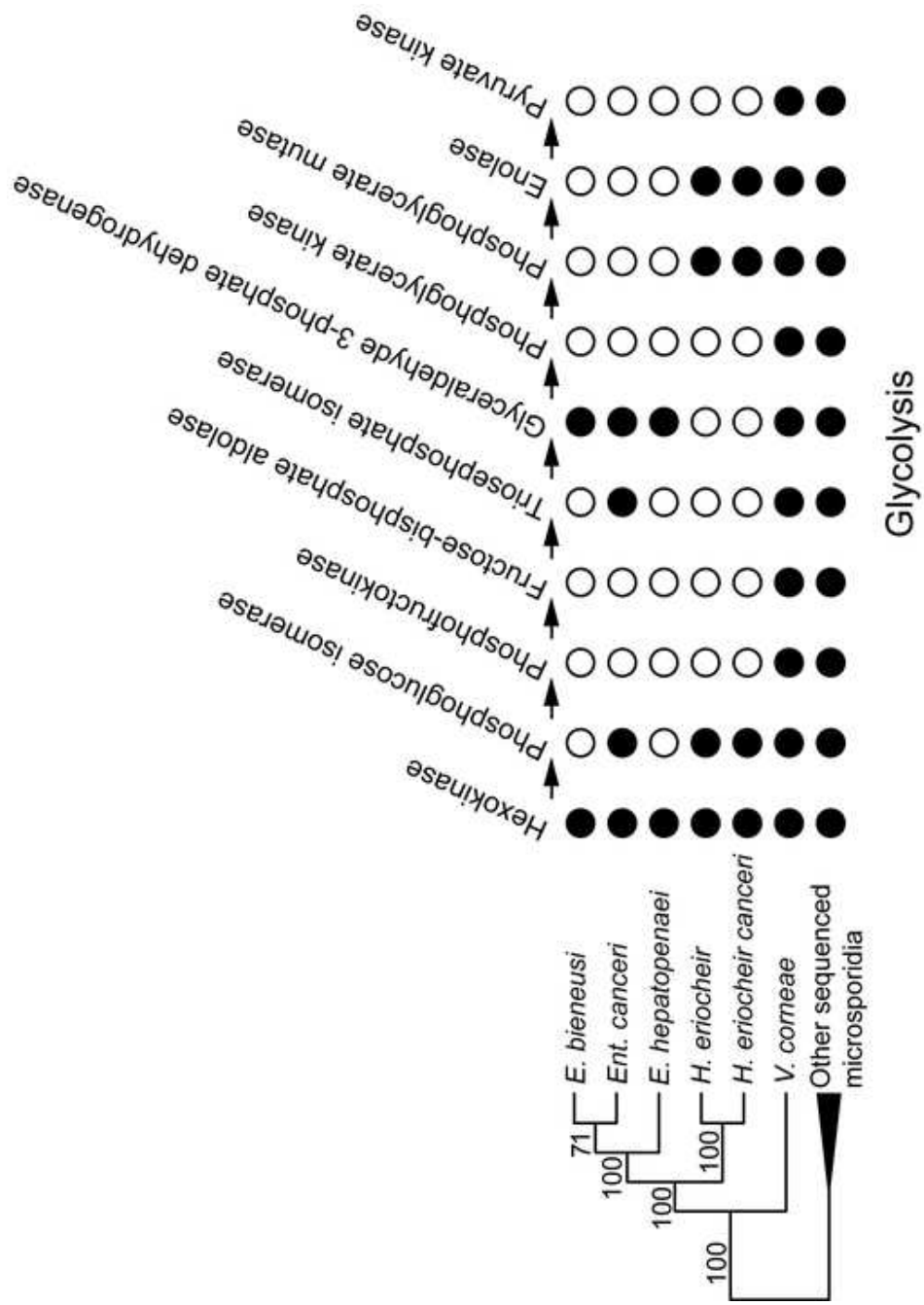
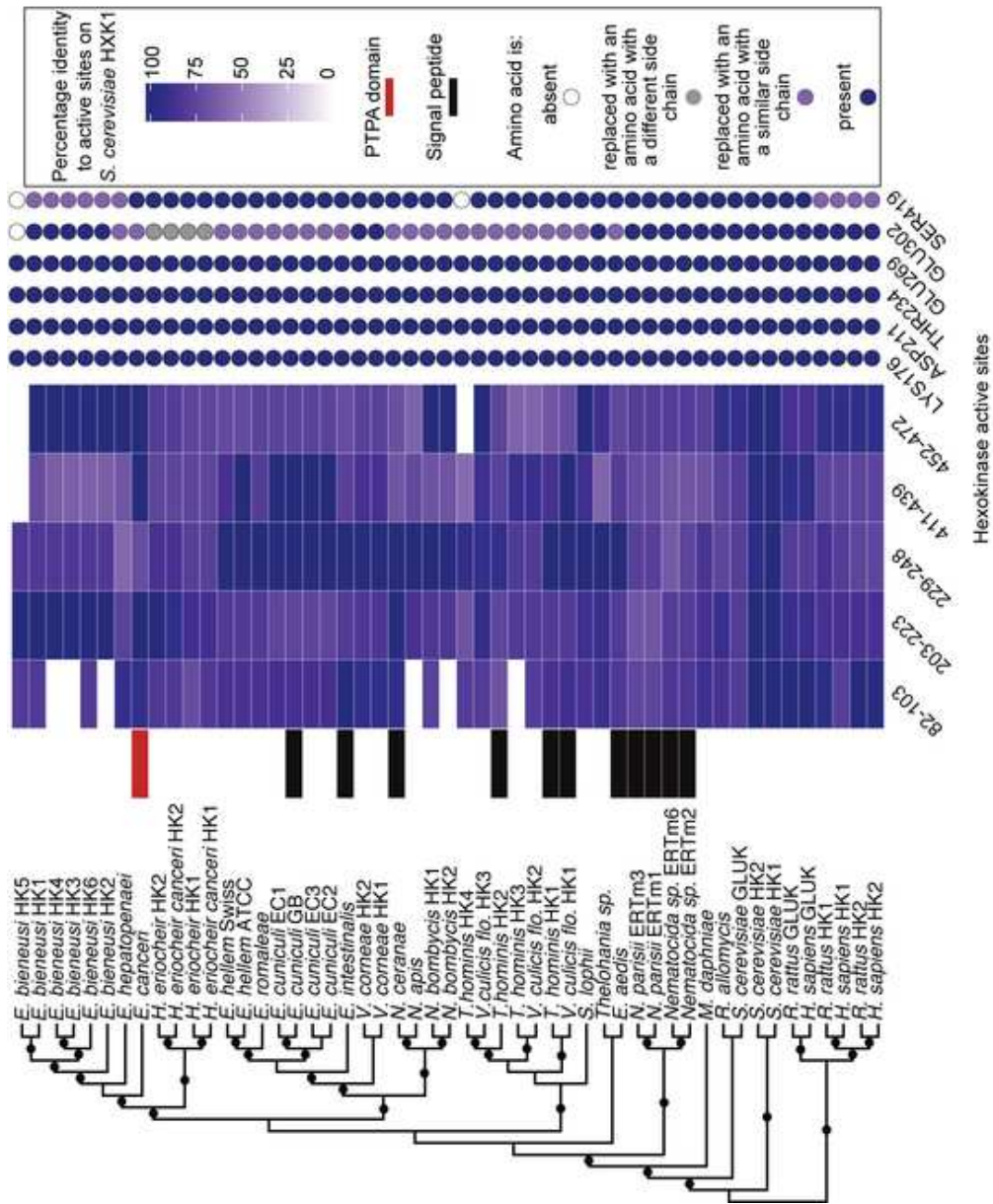


Figure 3

[Click here to download Figure Figure3.tif](#)



Appendix 14 Proteins exclusive to glycolytic genomes

Table 1: Protein clusters exclusive to glycolytic genomes

<i>Encephalitozoon cuniculi</i> gene ID	Gene ontology accession number	Annotation
ECU01_0370	GO:0005730	ribosome-associated chaperone zutin
ECU02_0230	GO:0005634	ISOPENTENYL-DIPHOSPHATE-DELTA-ISOMERASE II
ECU02_1370	GO:0004555	alpha-alpha-trehalase precursor
ECU03_0680	GO:0005945	6-phosphofructokinase
ECU03_0750	GO:0019013	U1 small nuclear ribonucleo a
ECU05_0320	GO:0004618	PHOSPHOGLYCERATE KINASE
ECU06_0490	GO:0005829	diphosphomevalonate decarboxylase
ECU06_0940	GO:0005777	3-ketoacyl thiolase
ECU06_1050	GO:0016021	hypothetical protein ECU06_1050
ECU07_1720	GO:0000166	Splicing factor 3B subunit 4
ECU08_0650	GO:0003899	DNA-directed RNA polymerase subunit E
ECU08_1530	GO:0003723	pseudouridylate synthase
ECU08_1620	GO:0005634	ser thr kinase
ECU10_0440	GO:0005634	ATP-DEPENDENT RNA HELICASE
ECU10_0510	GO:0004421	hydroxymethylglutaryl- synthase
ECU10_1510	GO:0005737	mevalonate kinase
ECU10_1590	GO:0005635	Choline-phosphate cytidylyltransferase partial
ECU10_1720	GO:0004420	3-HYDROXY-3-METHYLGLUTARYL REDUCTASE
ECU11_0230	GO:0004807	triosephosphate isomerase
ECU11_0620	GO:0016021	CDP-alcohol phosphatidyltransferase
ECU11_1810	GO:0005634	FARNESYL PYROPHOSPHATE SYNTHETASE
ECU09_0640	GO:0000287	pyruvate kinase
ECU09_0810	GO:0003924	Elongation factor G
ECU09_1260	GO:0005634	RIO kinase
ECU09_1500	GO:0005634	trna (uracil-5-)-methyltransferase trm9
ECU09_1780	GO:0005737	mevalonate kinase
ECU01_1070		hypothetical protein
ECU03_0570		hypothetical protein
ECU03_0880		hypothetical protein
ECU03_1020		hypothetical protein
ECU04_1520		hypothetical protein
ECU05_0830		hypothetical protein
ECU06_0970		hypothetical protein
ECU10_1550		hypothetical protein
ECU11_1410		hypothetical protein
ECU09_0420		hypothetical protein
ECU09_0650		hypothetical protein
ECU09_1280		hypothetical protein

Appendix 14: Phylogenomic analyses of 23 microsporidian species

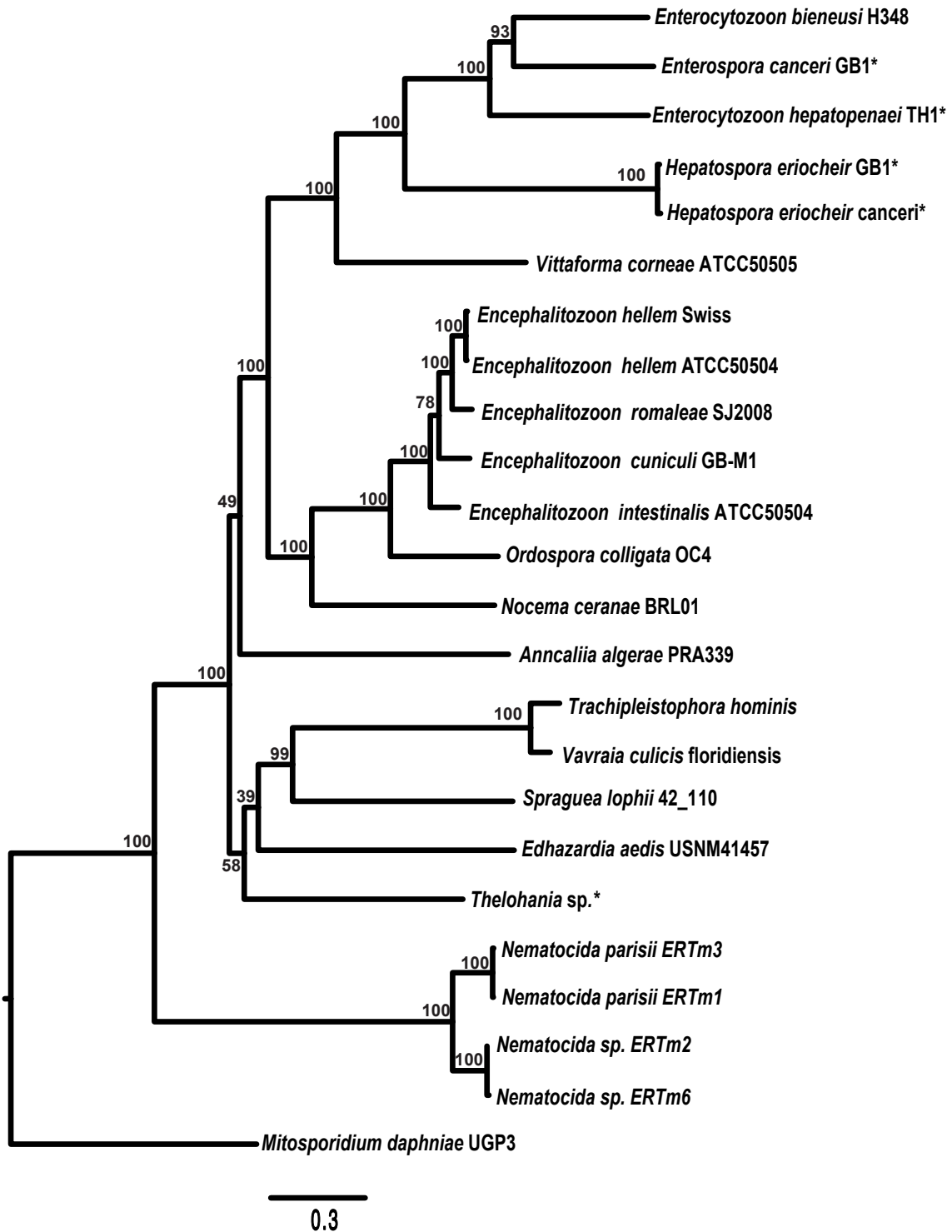


Figure 1: Phylogenetic assessment of 23 microsporidian species. Best-scoring maximum likelihood tree out of 100 bootstrapped trees. Values on nodes represent bootstrap support. Analyses were performed on a concatenated alignment of 21 proteins. Species with * are those whose genomic data were produced in this study. Branch lengths are proportional to substitution per site as indicated by the scale bar.

Bibliography

Abascal, F., Zardoya, R. & Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, 21(9), pp.2104-5.

ADFG, 2016. Commercial Salmon Catch, Effort & Value. *Commercial Fishing*.

Agersø, Y. & Petersen, A., 2007. The tetracycline resistance determinant Tet 39 and the sulphonamide resistance gene *sulII* are common among resistant *Acinetobacter* spp. isolated from integrated fish farms in Thailand. *The Journal of antimicrobial chemotherapy*, 59(1), pp.23-7.

Ahuatzi, D., Herrero, P., de la Cera, T. & Moreno, F., 2004. The glucose-regulated nuclear localization of hexokinase 2 in *Saccharomyces cerevisiae* is Mig1-dependent. *The Journal of biological chemistry*, 279(14), pp.14440-6.

Akashi, H. & Eyre-Walker, A., 1998. Translational selection and molecular evolution. *Current Opinion in Genetics & Development*, 8(6), pp.688-693.

Akiyoshi, D.E., Morrison, H.G., Lei, S., Feng, X., Zhang, Q., Corradi, N., Mayanja, H., Tumwine, J.K., Keeling, P.J., Weiss, L.M. & Tzipori, S., 2009. Genomic survey of the non-cultivable opportunistic human pathogen, *Enterocytozoon bieneusi*. *Public library of science pathogens*, 5(1).

Akman, L. & Aksoy, S., 2001. A novel application of gene arrays: *Escherichia coli* array provides insight into the biology of the obligate endosymbiont of tsetse flies. *Proceedings of the national academy of sciences of the United States of America*, 98(13), pp.7546-51.

Albig, W. & Entian, K.D., 1988. Structure of yeast glucokinase, a strongly diverged specific aldo-hexose-phosphorylating isoenzyme. *Gene*, 73(1), pp.141-52.

Alkan, C., Sajjadian, S. & Eichler, E.E., 2010. Limitations of next-generation genome sequence assembly. *Nature Methods*, 8(1), pp.61-65.

Amaro, F., Gilbert, J.A., Owens, S., Trimble, W. & Shuman, H.A., 2012. Whole-genome sequence of the human pathogen *Legionella pneumophila* serogroup 12 strain 570-CO-H. *Journal of bacteriology*, 194(6), pp.1613-4.

Andersson, J.O. & Andersson, S.G., 1999. Genome degradation is an

ongoing process in *Rickettsia*. *Molecular biology and evolution*, 16(9), pp.1178-91.

Andersson, S.G. & Kurland, C.G., 1995. Genomic evolution drives the evolution of the translation system. *Biochemistry and cell biology*, 73(11-12), pp.775-87.

Andersson, S.G. & Kurland, C.G., 1998. Reductive evolution of resident genomes. *Trends in microbiology*, 6(7), pp.263-8.

Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Pontén, T., Alsmark, U.C., Podowski, R.M., Näslund, A.K., Eriksson, A.S., Winkler, H.H. & Kurland, C.G., 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, 396(6707), pp.133-40.

Angellotti, M.C., Bhuiyan, S.B., Chen, G. & Wan X.-F., 2007. CodonO: codon usage bias analysis within and across genomes. *Nucleic acids research*, 35, pp.W132-6.

Aurrecoechea, C., Barreto, A., Brestelli, J., Brunk, B.P., Caler, E.V., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Iodice, J., Kissinger, J.C., Kraemer, E.T., Li, W., Nayak, V., Pennington, C., Pinney, D.F., Pitts, B., Roos, D.S., Srinivasamoorthy, G., Stoeckert, C.J.Jr., Treatman, C. & Wang, H., 2011. AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. *Nucleic acids research*, 39, pp.D612-9.

Ausubel, F. M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith J.A. & Struhl, K., 2002. *Short protocols in molecular biology*. , 2, p.800.

Badil, S., Elliott, D.G., Kurobe, T., Hedrick, R.P., Clemens, K., Blair, M. & Purcell, M.K., 2011. Comparative evaluation of molecular diagnostic tests for *Nucleospora salmonis* and prevalence in migrating juvenile salmonids from the Snake River, USA. *Journal of aquatic animal health*, 23(1), pp.19-29.

Baker, M.D., Vossbrinck, C.R., Maddox, J.V. & Undeen A.H., 1994. Phylogenetic relationships among *Vairimorpha* and *Nosema* species (Microspora) based on ribosomal RNA sequence data. *Journal of invertebrate pathology*, 64(2), pp.100-6.

- Bakken, L. & Olsen, R., 1989. DNA-content of soil bacteria of different cell size. *Soil Biology and Biochemistry*, 21(6), pp.789-793.
- Balbani, G., 1882. Sur les microsporidies ou psorospermies des articles. *Comptes rendus de l'académie des sciences.*, 95, pp.1168-1171.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A. & Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5), pp.455-77.
- Bano, N., Romano, J.D., Jayabalasingham, B. & Coppens, I., 2007. Cellular interactions of *Plasmodium* liver stage with its host mammalian cell. *International Journal for Parasitology*, 37(12), pp.1329-41.
- Bao, W., Kojima, K.K. & Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1), p.11.
- Bartholomeu, D.C., Cerqueira, G.C., Leão¹, A.C.A., daRocha, W.D., Pais, F.S., Macedo, C., Djikeng, A., Teixeira, S.M.R. & El-Sayed, N.M., 2009. Genomic organization and expression profile of the mucin-associated surface protein (masp) family of the human pathogen *Trypanosoma cruzi*. *Nucleic acids research*, 37(10), pp.3407-3417.
- Bateman, K.S., Hicks, R.J. & Stentiford, G.D., 2011. Disease profiles differ between non-fished and fished populations of edible crab (*Cancer pagurus*) from a major commercial fishery. *ICES Journal of Marine Science*, 68(10), pp.2044-2052.
- Becker, C. & Türkay, M., 2010. Taxonomy and morphology of European pea crabs (Crustacea: Brachyura: Pinnotheridae). *Journal of Natural History*, 44(25-26), pp.1555-1575.
- Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R. & Boutell, J.M., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), pp.53-9.

Berg, J.M., Tymoczko, J.L. & Stryer, L., 2006. *Biochemistry* 6th ed., New York: W. H. Freeman and Company.

Berke, S.K., Miller, M. & Woodin, A.S., 2006. Modelling the energy-mortality trade-offs of invertebrate decorating behaviour. *Evolutionary ecology research*, 8, pp.1409-1425.

Bianchi, C., Genova, M.L., Parenti Castelli, G. & Lenaz, G., 2004. The Mitochondrial Respiratory Chain Is Partially Organized in a Supercomplex Assembly: Kinetic evidence using flux control analysis. *Journal of Biological Chemistry*, 279(35), pp.36562–9.

Bigliardi, E. & Sacchi, L., 2001. Cell biology and invasion of the microsporidia. *Microbes and infection*, 3(5), pp.373-379.

Bigliardi, E., Gatti, S. & Sacchi, L., 1997. Ultrastructure of microsporidian spore wall: The *Encephalitozoon cuniculi* exospore. *Italian journal of zoology*, 64(1), pp.1-5.

Bigliardi, E., Selmi, M.G., Lupetti, P., Corona, S., Gatti, S., Scaglia, M. & Sacchi, L., 1996. Microsporidian spore wall: ultrastructural findings on *Encephalitozoon hellem* exospore. *The Journal of eukaryotic microbiology*, 43(3), pp.181-6.

Bininda-Emonds, O.R.P., 2004. The evolution of supertrees. *Trends in Ecology & Evolution*, 19(6), pp.315-322.

Bininda-Emonds, O.R.P., Gittleman, J.L. & Steel, M.A., 2002. The (Super)Tree of Life: Procedures, Problems, and Prospects. *Annual Review of ecology and systematics*, 33, pp.265-289.

Blin, C., Panserat, S., Médale, F., Gomes, E., Breque, J., Kaushik, S. & Krishnamoorthy, R., 1999. Teleost liver hexokinase- and glucokinase-like enzymes: partial cDNA cloning and phylogenetic studies in rainbow trout (*Oncorhynchus mykiss*), common carp (*Cyprinus carpio*) and gilthead seabream (*Sparus aurata*). *Fish physiology and biochemistry*, 21(2), pp.93-102.

Boisvert, S., Laviolette, F. & Corbeil, J., 2010. Ray: simultaneous assembly of

reads from a mix of high-throughput sequencing technologies. *Journal of computational biology*, 17(11), pp.1519-33.

Bolger, A.M., Lohse, M. & Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), pp.2114-20.

Bonner, W.M., 1975. Protein migration into nuclei. I. Frog oocyte nuclei in vivo accumulate microinjected histones, allow entry to small proteins, and exclude large proteins. *The Journal of Cell Biology*, 64(2), pp.421-30.

Boon, J.P., Lewis, W.E., Tjoen-A-Choy, M.R., Allchin, C.R., Law, R.J., de Boer, J., ten Hallers-Tjabbes, C.C. & Zegers, B.N., 2002. Levels of polybrominated diphenyl ether (PBDE) flame retardants in animals representing different trophic levels of the North Sea food web. *Environmental science & technology*, 36(19), pp.4025-4032.

Borgström, E., Lundin, S. & Lundeberg, J., 2011. Large scale library generation for high throughput sequencing. *Public library of science one*, 6(4).

Bork, P., Sander, C. & Valencia, A., 1993. Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein science*, 2(1), pp.31-40.

Boyer, P.D., 2000. Catalytic site forms and controls in ATP synthase catalysis. *Biochimica et biophysica acta*, 1458(2-3), pp.252-62.

Brachmann, C.B., Davies, A., Cost, G.J., Caputo, E., Li, J., Hieter, P. & Boeke, J.D. 1998. Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast*, 14(2), pp.115-32.

Bravo, S., 1996. *Enterocytozoon salmonis* in Chile. *American fisheries society, fish health section newsletter*, 24, pp.12-13.

Breton, J., Bart-Delabesse, E., Biligui, S., Carbone, A., Seiller, X., Okome-Nkoumou, M., Nzamba, C., Kombila, M., Accoceberry, I. & Thellier, M., 2007. New highly divergent rRNA sequence among biodiverse genotypes of *Enterocytozoon bieneusi* strains isolated from humans in Gabon and Cameroon. *Journal of clinical microbiology*, 45(8), pp.2580-9.

Briffa, M. & Elwood, R.W., 2005. Metabolic consequences of shell choice in *Pagurus bernhardus*: do hermit crabs prefer cryptic or portable shells? *Behavioral ecology and sociobiology*, 59(1), pp.143-148.

Briffa, M., Bridger, D. & Biro, P.A., 2013. How does temperature affect behaviour? Multilevel analysis of plasticity, personality and predictability in hermit crabs. *Animal behaviour*, 86(1), pp.47-54.

Brown, A.M. V, Kent, M.L. & Adamson, M.L., 2010. Description of five new *Loma* (Microsporidia) species in pacific fishes with redesignation of the type species *Loma morhua* Morrison & Sprague, 1981, based on morphological and molecular species-boundaries tests. *The Journal of eukaryotic microbiology*, 57(6), pp.529-53.

Brown, A.M.V & Adamson, M.L., 2006. Phylogenetic distance of *Thelohania butleri* Johnston, Vernick, and Sprague, 1978 (Microsporidia; Thelohaniidae), a parasite of the smooth pink shrimp *Pandalus jordani*, from its congeners suggests need for major revision of the genus *Thelohenia* Henneguy, 1982. *The Journal of eukaryotic microbiology*, 53(6), pp.445-455.

Brown, C.J., Todd, K.M. & Rosenzweig, R.F., 1998. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Molecular biology and evolution*, 15(8), pp.931-42.

Brown, J.R. & Doolittle, W.F., 1999. Gene descent, duplication, and horizontal transfer in the evolution of glutamyl- and glutaminyl-tRNA synthetases. *Journal of molecular evolution*, 49(4), pp.485-495.

Bui, E.T., Bradley, P.J. & Johnson, P.J., 1996. A common evolutionary origin for mitochondria and hydrogenosomes. *Proceedings of the National academy of sciences of the United States of America*, 93(18), pp.9651-6.

Cáceres, A.J., Quiñones, W., Gualdrón, M., Cordeiro, A., Avilán, L., Michels, P.A. & Concepción, J.L., 2007. Molecular and biochemical characterization of novel glucokinases from *Trypanosoma cruzi* and *Leishmania* spp. *Molecular and biochemical parasitology*, 156(2), pp.235-45.

Cali, A., Kotler, D.P. & Orenstein, J.M., 1993. *Septata intestinalis* n.g., n. sp., an intestinal microsporidian associated with chronic diarrhea and

dissemination in AIDS patients. *The Journal of eukaryotic microbiology*, 40(1), pp.101-112.

Cali, A., Weiss, L.M. & Takvorian, P.M., 2002. *Brachiola algerae* spore membrane systems, their activity during extrusion, and a new structural entity, the multilayered interlaced network, associated with the polar tube and the sporoplasm. *The Journal of eukaryotic microbiology*, 49(2), pp.164-74.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L., 2009. BLAST+: architecture and applications. *BioMed central bioinformatics*, 10, p.421.

Campbell, S.E., Williams, T.A., Yousuf, A., Soanes, D.M., Paszkiewicz, K.H. & Williams, B.A.P., 2013. The genome of *Spraguea lophii* and the basis of host-microsporidian interactions. *Public library of science genetics*, 9(8).

Canback, B., Andersson, S.G.E. & Kurland, C.G., 2002. The global phylogeny of glycolytic enzymes. *Proceedings of the national academy of sciences of the United States of America*, 99(9), pp.6097-102.

Canning, E.U. & Hazard, E.I., 1982. Genus *Pleistophora* Gurley, 1893: An assemblage of at least three genera. *The Journal of protozoology*, 29(1), pp.39-49.

Canning, E.U. & Hollister, W.S., 1990. *Enterocytozoon bieneusi* (Microspora): prevalence and pathogenicity in AIDS patients. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 84(2), pp.181-186.

Canning, E.U. & Hollister, W.S., 1992. Human infections with microsporidia. *Reviews in Medical Microbiology*, 3, pp.35-42.

Canning, E.U. & Vavra, J., 2000. Phylum microsporidia balbiani, 1982. In J. Lee, G. F. Leedale, & P. Bradbury, eds. *The Illustrated Guide to the Protozoa*. Society of Protozoologists.

Canning, E.U., 1953. A new microsporidian, *Nosema locustae* n.sp., from the fat body of the African migratory locust, *Locusta migratoria migratorioides* R. & F. *Parasitology*, 43(3-4), pp.287-290.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C.,

Alvarado, A.S. & Yandell, M., 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research*, 18(1), pp.188-96.

Capella-gutiérrez, S., Marcet-houben, M. & Gabaldón, T., 2012. Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. *BioMed central biology*, 10(1), p.1.

Capella-Gutierrez, S., Silla-Martinez, J.M. & Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15), pp.1972-1973.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M. & Gormley, N., 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME journal*, 6(8), pp.1621-4.

Cárdenas, M.L., Cornish-Bowden, A. and Ureta, T., 1998. Evolution and regulatory role of the hexokinases. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1401(3), pp.242-264.

Carver, T., Berriman, M., Tivey, A., Patel, C., Böhme, U., Barrell, B.G., Parkhill, J. & Rajandream, M.A., 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, 24(23), pp.2672-6.

Casadevall, A., 2008. Evolution of intracellular pathogens. *Annual review of microbiology*, 62, pp.19-33.

Cascianelli, G., Villani, M., Tosti, M., Marini, F., Bartoccini, E., Magni, M.V. & Albi, E., 2008. Lipid microdomains in cell nucleus. *Molecular biology of the cell*, 19(12), pp.5289–95.

Cavalier-Smith, T., 1983. A 6-kingdom classification and a unified phylogeny. *Endocytobiology II*, pp.1027-1034.

Cavalier-Smith, T., 1987. The origin of cells: a symbiosis between genes, catalysts, and membranes. *Cold Spring Harbor symposia on quantitative*

biology, 52, pp.805-24.

Cavalier-Smith, T., 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Annals of botany*, 95(1), pp.147-75.

Chacin-Bonilla, L., Panunzio, A.P., Monsalve-Castillo, F.M., Parra-Cepeda, I.E. & Martinez, R., 2006. Microsporidiosis in Venezuela: prevalence of intestinal microsporidiosis and its contribution to diarrhea in a group of human immunodeficiency virus-infected patients from Zulia State. *The American journal of tropical medicine and hygiene*, 74(3), pp.482-6.

Chaisson, M., Pevzner, P. & Tang, H., 2004. Fragment assembly with short reads. *Bioinformatics*, 20(13), pp.2067-74.

Chakravarty, A.K. & Shuman, S., 2011. RNA 3'-Phosphate Cyclase (RtcA) Catalyzes Ligase-like Adenylation of DNA and RNA 5'-Monophosphate Ends. *Journal of biological chemistry*, 286(6), pp.4117-4122.

Chalifoux, L.V., MacKey, J., Carville, A., Shvetz, D., Lin, K.C., Lackner, A. & Mansfield, K.G., 1998. Ultrastructural morphology of *Enterocytozoon bieneusi* in biliary epithelium of rhesus macaques (*Macaca mulatta*). *Veterinary pathology*, 35(4), pp.292-6.

Chan, P.P. & Lowe, T.M., 2009. GtRNADB: a database of transfer RNA genes detected in genomic sequence. *Nucleic acids research*, 37, pp.D93-7.

Charles, H. & Ishikawa, H., 1999. Physical and genetic Mmap of the genome of *Buchnera*, the primary endosymbiont of the pea aphid *Acyrtosiphon pisum*. *Journal of molecular evolution*, 48(2), pp.142-150.

Chayaburakul, K., Nash, G., Pratanpipat, P., Sriurairatana, S. & Withyachumnarnkul, B., 2004. Multiple pathogens found in growth-retarded black tiger shrimp *Penaeus monodon* cultivated in Thailand. *Diseases of aquatic organisms*, 60(2), pp.89-96.

Chen, K., Chen, L., Fan, X., Wallis, J., Ding, L. & Weinstock, G., 2014. TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome research*, 24(2), pp.310-7.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L. & McAdams, H.H., 2004. Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the national academy of sciences of the United States of America*, 101(10), pp.3480-5.

Chen, T., Ning, D., Sun, H., Li, R., Shang, M., Li, X., Wang, X., Chen, W., Liang, C., Li, W. & Mao, Q., 2014. Sequence Analysis and Molecular Characterization of *Clonorchis sinensis* Hexokinase, an Unusual Trimeric 50-kDa Glucose-6-Phosphate-Sensitive Allosteric Enzyme. *Public library of science one*, 9(9).

Chen, Y. ping, Pettis, J.S., Zhao, Y., Liu, X., Tallon, L.J., Sadzewicz, L.D., Li, R., Zheng, H., Huang, S., Zhang, X. & Hamilton, M.C., 2013. Genome sequencing and comparative genomics of honey bee microsporidia, *Nosema apis* reveal novel insights into host-parasite interactions. *BioMed central genomics*, 14(1), p.451.

Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R. & Fisk, D.G., 2012. *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic acids research*, 40, pp.D700-5.

Chilmonczyk, S., Cox, W.T. & Hedrick, R.P., 1991. *Enterocytozoon salmonis* n. sp.: an intranuclear microsporidium from salmonid fish. *The Journal of protozoology*, 38(3), pp.264-9.

Chisholm, S.T., Coaker, G., Day, B. & Staskawicz, B.J., 2006. Host-Microbe Interactions: Shaping the Evolution of the Plant Immune Response. *Cell*, 124(4), pp.803-814.

Cho, Y.-H., Yoo, S.-D. & Sheen, J., 2006. Regulatory functions of nuclear hexokinase1 complex in glucose signaling. *Cell*, 127(3), pp.579-89.

Claeyssen, É., Wally, O., Matton, D.P., Morse, D. & Rivoal, J., 2006. Cloning, expression, purification, and properties of a putative plasma membrane hexokinase from *Solanum chacoense*. *Protein expression and purification*, 47(1), pp.329-39.

Clark, C.G. & Roger, A.J., 1995. Direct evidence for secondary loss of

mitochondria in *Entamoeba histolytica*. *Proceedings of the national academy of sciences of the United States of America*, 92(14), pp.6518-21.

Clark, P.F., Rainbow, P.S., Robbins, R.S., Smith, B., Yeomans, W.E., Thomas, M. & Dobson, G., 1998. The alien Chinese mitten crab, *Eriocheir sinensis* (Crustacea: Decapoda: Brachyura), in the Thames catchment. *Journal of the Marine Biological Association of the United Kingdom*, 78(4), pp.1215-1221.

Clements, A. & Paterson, G., 1981. The analysis of mortality and survival rates in wild populations of mosquitoes. *Journal of applied ecology*.

Coil, D., Jospin, G. & Darling, A.E., 2015. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics*, 31(4), pp.587-9.

Colloca, F. & Cerasi, S., 2005. Cultured Aquatic Species Information Programme. *Sparus aurata*. Cultured Aquatic Species Information Programme. *FAO Fisheries and Aquaculture Department*. Available at: http://www.fao.org/fishery/culturedspecies/Sparus_aurata/en#tcNA00EA [Accessed January 31, 2016].

Compeau, P.E.C., Pevzner, P.A. & Tesler, G., 2011. How to apply de Bruijn graphs to genome assembly. *Nature biotechnology*, 29(11), pp.987-91.

Conant, G.C. & Wagner, A., 2002. GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic acids research*, 30(15), pp.3378-86.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. & Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18), pp.3674-6.

Corliss, J.O., 1994. An interim utilitarian ('User-friendly') hierarchical classification and characterization of the protists. *Acta Protozoologica*.

Cornman, R.S., Chen, Y.P., Schatz, M.C., Street, C., Zhao, Y., Desany, B., Egholm, M., Hutchison, S., Pettis, J.S., Lipkin, W.I. & Evans, J.D., 2009. Genomic analyses of the microsporidian *Nosema ceranae*, an emergent

- pathogen of honey bees. *Public library of science pathogens*, 5(6).
- Corradi, N. & Keeling, P.J., 2009. Microsporidia: a journey through radical taxonomical revisions. *Fungal Biology Reviews*, 23(1-2), pp.1-8.
- Corradi, N., Akiyoshi, D.E., Morrison, H.G., Feng, X., Weiss, L.M., Tzipori, S., & Keeling, P.J., 2007. Patterns of genome evolution among the microsporidian parasites *Encephalitozoon cuniculi*, *Antonospora locustae* and *Enterocytozoon bieneusi*. *Public library of science one*, 2(12).
- Corradi, N., Haag, K. L., Pombert, J.-F., Ebert, D., & Keeling, P. J., 2009. Draft genome sequence of the *Daphnia* pathogen *Octospora bayeri*: insights into the gene content of a large microsporidian genome and a model for host-parasite interactions. *Genome biology*, 10(10).
- Corradi, N., Pombert, J.-F., Farinelli, L., Didier, E.S., & Keeling, P.J., 2010. The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*. *Nature communications*, 1, p.77.
- Cortés, A., 2005. A chimeric *Plasmodium falciparum* Pfnbp2b/Pfnbp2a gene originated during asexual growth. *International Journal for Parasitology*, 35(2),pp.125-30
- Cuomo, A.C., Desjardins, C.A., Bakowski, M.A., Goldberg, J., Ma, A.T., Becnel, J.J. Didier, E.S., Fan, L., Heiman, D.I., Levin, J.Z., Young, S., Zeng, Q. & Troemel, E.R. 2012. Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth. *Genome research*, 22(12), pp.2478-88.
- Cupples, C.G. & Miller, J.H., 1988. Effects of amino acid substitutions at the active site in *Escherichia coli* beta-galactosidase. *Genetics*, 120(3), pp.637-44.
- Daboussi, M.-J. & Capy, P., 2003. Transposable elements in filamentous fungi. *Annual review of microbiology*, 57, pp.275-99.
- Das, R., Hegyi, H. & Gerstein, M., 2000. Genome analyses of spirochetes: a study of the protein structures, functions and metabolic pathways in *Treponema pallidum* and *Borrelia burgdorferi*. *Journal of molecular*

microbiology and biotechnology, 2(4), pp.387-92.

Davis, J.A. & Freeze, H.H., 2001. Studies of mannose metabolism and effects of long-term mannose ingestion in the mouse. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1528(2), pp.116-126.

Dean, P., Major P., Nakjang S., Hirt R.P. & Embley T.M., 2014. Transport proteins of parasitic protists and their role in nutrient salvage. *Frontiers in plant science*, 5, p.153.

DeAngelis, Y.M., Saunders, C.W., Johnstone, K.R., Reeder, N.L., Coleman, C.G., Kaczvinsky, J.R. Jr, Gale, C., Walter, R., Mekel, M., Lacey, M.P., Keough, T.W., Fieno, A., Grant, R.A., Begley, B, Sun, Y, Fuentes, G., Youngquist, R.S., Xu, J. & Dawson, T.L. Jr., 2007. Isolation and expression of a *Malassezia globosa* lipase gene, LIP1. *The Journal of investigative dermatology*, 127(9), pp.2138-46.

Decraene, V., Lebbad, M., Botero-kleiven, S., Gustavsson, A.-M., & Löfdahl, M., 2012. First reported foodborne outbreak associated with microsporidia, Sweden, October 2009. *Epidemiology and infection*, 140(3), pp.519-27.

del Aguila, C., Izquierdo, F., Navajas, R., Pieniazek, N.J., Miró, G., Alonso, A.I., Da Silva, A.J. & Fenoy, S., 1999. *Enterocytozoon bieneusi* in animals: rabbits and dogs as new hosts. *Journal of Eukaryotic Microbiology*, 46(5), pp.8S-9S.

Delsuc, F., Brinkmann, H. & Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. *Nature reviews genetics*, 6(5), pp.361-75.

Dengjel, B., Zahler, M., Hermanns, W., Heinritzi, K., Spillmann, T., Thomschke, A., Löscher, T., Gothe, R. & Rinder, H., 2001. Zoonotic potential of *Enterocytozoon bieneusi*. *Journal of clinical microbiology*, 39(12), pp.4495-9.

Desjardins, C.A., Sanscrainte, N.D., Goldberg, J.M., Heiman, D., Young, S., Zeng, Q., Madhani, H.D., Becnel, J.J. & Cuomo, C.A., 2015. Contrasting host-pathogen interactions and genome evolution in two generalist and specialist microsporidian pathogens of mosquitoes. *Nature communications*, 6, p.7121.

Desportes-livage, I., Chilmonczyk, S., Hedrick, R., Ombrouck, C., Monge, D., Maiga, I. & Gentilini, M., 1996. Comparative development of two microsporidiana species: *Enterocytozoon bieneusi* and *Enterocytozoon salmonis*, reported in AIDS patients and salmonid fish, respectively. *The Journal of eukaryotic microbiology*, 43(1), pp.49-60.

Desportes-Livage, I., I. Hilmarsdottir, C. Romana, S. Tanguy, A. Datry, & M. Gentilini., 1991. Characteristics of the microsporidian *Enterocytozoon bieneusi*: a consequence of its development within short-living enterocytes. *Journal of Protozoology*, 32, p.111S-113S.

Desportes, I., Charpentier, Y.L., Galian, A., Bernard, F., Cochand-Priollet, B., Lavergne, A., Ravisse, P. & Modigliani, R., 1985. Occurrence of a new microsporidan: *Enterocytozoon bieneusi* n.g., n. sp., in the enterocytes of a human patient with AIDS. *The Journal of protozoology*, 32(2), pp.250-4.

Didier, E.S., Visvesvara, G.S., Baker, M.D., Rogers, L.B., Bertucci, D.C., De Groote, M.A. & Vossbrinck, C.R., 1996. A microsporidian isolated from an AIDS patient corresponds to *Encephalitozoon cuniculi* III, originally isolated from domestic dogs. *Journal of Clinical Microbiology*, 34(11), pp.2835-2837.

Ding, Z., Meng, Q., Liu, H., Yuan, S., Zhang, F., Sun, M., Zhao, Y., Shen, M., Zhou, G., Pan, J. and Xue, H., 2016. First case of hepatopancreatic necrosis disease in pond-reared Chinese mitten crab, *Eriocheir sinensis*, associated with microsporidian. *Journal of fish diseases*, 39(9), pp.1043-51.

Docker, M.F., Kent, M.L., Hervio, D.M., Khattra, J.S., Weiss, L.M., Cali, A.N.N. & Devlin, R.H., 1997. Ribosomal DNA sequence of *Nucleospora salmonis* Hedrick, Groff and Baxa, 1991 (Microsporea: Enterocytozoonidae): Implications for phylogeny and nomenclature. *The Journal of eukaryotic microbiology*, 44(1), pp.55-60.

Doehlert, D.C., Kuo, T.M. & Felker, F.C., 1988. Enzymes of sucrose and hexose metabolism in developing kernels of two inbreds of maize. *Plant physiology*, 86(4), pp.1013-9.

Doimi, M., 1996. A new winter disease in sea bream (*Sparus aurata*): a preliminary report. *Bulletin of the European Association of Fish Pathologists*, 353

16(1), pp.17-18.

Dolgikh, V.V., Senderskiy, I.V., Pavlova, O.A., Naumov, A.M. & Beznoussenko, G.V., 2011. Immunolocalization of an alternative respiratory chain in *Antonospora (Paranosema) locustae* spores: mitochondria retain their role in microsporidial energy metabolism. *Eukaryotic cell*, 10(4), pp.588-93.

Doménech, A., Fernández-Garayzábal, J.F., García, J.A., Cutuli, M.T., Blanco, M., Gibello, A., Moreno, M.A. & Domínguez, L., 1999. Association of *Pseudomonas anguilliseptica* infection with “winter disease” in sea bream, *Sparus aurata* L. *Journal of Fish Diseases*, 22(1), pp.69-71.

Dover, G. & Coen, E., 1981. Springcleaning ribosomal DNA: a model for multigene evolution? *Nature*, 290(5809), pp.731-732.

Draculic, T., Dawes, I.W. & Grant, C.M., 2000. A single glutaredoxin or thioredoxin gene is essential for viability in the yeast *Saccharomyces cerevisiae*. *Molecular microbiology*, 36(5), pp.1167-74.

Duarte, C.M., Holmer, M., Olsen, Y., Soto, D., Marbà, N., Guiu, J., Black, K. & Karakassis, I., 2009. Will the oceans help feed humanity? *BioScience*, 59(11), pp.967-976.

Duret, L., 2002. Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, 12(6), pp.640-649.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5), pp.1792-7.

Edlind, T., Visvesvara, G., Li, J. & Katiyar, S., 1996. *Cryptosporidium* and microsporidial beta-tubulin sequences: predictions of benzimidazole sensitivity and phylogeny. *The Journal of eukaryotic microbiology*, 41(5), p.38S.

Eichberg, J., Whittaker, V.P., Dawson, R.M. & Dawson, R.M.C., 1964. Distribution of lipids in subcellular particles of guinea-pig brain. *The Biochemical journal*, 92(1), pp.91–100.

Elston, R.A., Kent, M.L. & Harrell, L.H., 1987. An intranuclear microsporidium

associated with acute anemia in the chinook salmon, *Oncorhynchus tshawytscha*. *The Journal of protozoology*, 34(3), pp.274-7.

Elwood, R.W. & Appel, M., 2009. Pain experience in hermit crabs? *Animal Behaviour*, 77(5), pp.1243-1246.

Engel, P. & Moran, N.A., 2013. The gut microbiota of insects - diversity in structure and function. *Federation of European microbiological societies microbiology reviews*, 37(5), pp.699-735.

Entian, K.D. & Fröhlich, K.U., 1984. *Saccharomyces cerevisiae* mutants provide evidence of hexokinase PII as a bifunctional enzyme with catalytic and regulatory domains for triggering carbon catabolite repression. *Journal of bacteriology*, 158(1), pp.29-35.

Entian, K.D., Kopetzki, E., Fröhlich, K.U. & Mecke, D., 1984. Cloning of hexokinase isoenzyme PI from *Saccharomyces cerevisiae*: PI transformants confirm the unique role of hexokinase isoenzyme PII for glucose repression in yeasts. *Molecular & general genetics : MGG*, 198(2), pp.50-4.

European Commission, 2015. Gilthead seabream. Available at: http://ec.europa.eu/fisheries/marine_species/farmed_fish_and_shellfish/sea_bream/index_en.htm.

FAO, 2002. World production and markets for lumpfish eggs and lumpfish caviar. Available at: <http://www.fao.org/3/a-a0685e/a0685e02.pdf>.

Fast, N.M. & Keeling, P.J., 2001. Alpha and beta subunits of pyruvate dehydrogenase E1 from the microsporidian *Nosema locustae*: mitochondrion-derived carbon metabolism in microsporidia. *Molecular and biochemical parasitology*, 117(2), pp.201-209.

Fast, N.M., Law, J.S., Williams, B.A.P & Keeling, P.J., 2003. Bacterial catalase in the microsporidian *Nosema locustae*: Implications for microsporidian metabolism and genome evolution. *Eukaryotic cell*, 2(5), pp.1069-1075.

Fast, N.M., Logsdon, J.M. & Doolittle, W.F., 1999. Phylogenetic analysis of the TATA box binding protein (TBP) gene from *Nosema locustae*: evidence

for a microsporidia-fungi relationship and spliceosomal intron loss. *Molecular biology and evolution*, 16(10), pp.1415-1419.

Fayer, R. & Santin-Duran, M., 2014. Epidemiology of Microsporidia in human infections, In L. M. Weiss & J. J. Becnel, eds. *Microsporidia: pathogens of opportunity*. Pondicherry: John Wiley and Sons, inc., pp. 135-164.

Feldmeyer, B., Wheat, C.W., Krezdorn, N., Rotter, B. & Pfenninger, M., 2011. Short read Illumina data for the *de novo* assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance. *BioMed central genomics*, 12, p.317.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4), pp.401-410.

Feng, Y., Li, N., Dearen, T., Lobo, M.L., Matos, O., Cama, V. & Xiao, L., 2011. Development of a multilocus sequence typing tool for high-resolution genotyping of *Enterocytozoon bieneusi*. *Applied and environmental microbiology*, 77(14), pp.4822-8.

Findley, A.M., Weidner, E.H., Carman, K.R., Xu, Z. & Godbar, J.S., 2005. Role of the posterior vacuole in *Spraguea lophii* (Microsporidia) spore hatching. *Folia parasitologica*, 52(1-2), pp.111-7.

Fisher, D.J., Fernández, R.E. & Maurelli, A.T., 2013. *Chlamydia trachomatis* transports NAD via the Npt1 ATP/ADP translocase. *Journal of bacteriology*, 195(15), pp.3381-6.

Fisheries Research Services (FRS) Marine Laboratory. 2016. Code of Practice to Avoid the Introduction of *Gyrodactylus salaris* to GB. [ONLINE] Available at: <http://www.gov.scot/uploads/documents/copgyrod.pdf>. [Accessed 17 September 2016].

Flegel, T.W., 2012. Historic emergence, impact and current status of shrimp pathogens in Asia. *Journal of invertebrate pathology*, 110(2), pp.166-73.

Foltz, J.R., Plant, K.P., Overturf, K., Clemens, K. & Powell, M.S., 2009. Detection of *Nucleospora salmonis* in steelhead trout, *Oncorhynchus mykiss* (Walbaum), using quantitative polymerase chain reaction (qPCR). *Journal of*

fish diseases, 32(6), pp.551-5.

Foxx, J. & Siddall, M.E., 2015. The Road To Cnidaria: History of Phylogeny of the Myxozoa. *The Journal of parasitology*, 101(3), pp.269-74.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L. & Postlethwait, J., 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4), pp.1531-45.

Fothergill-Gilmore, L.A. & Michels, P.A.M., 1993. Evolution of glycolysis. *Progress in biophysics and molecular biology*, 59(2), pp.105-235.

Fournier, S., Liguory, O., Santillana-Hayat, M., Guillot, E., Sarfati, C., Dumoutier, N., Molina, J.M. & Derouin, F., 2000. Detection of microsporidia in surface water: a one-year follow-up study. *Federation of European microbiological societies immunology and medical microbiology*, 29(2), pp.95-100.

Fraser-Liggett, C.M., 2005. Insights on biology and evolution from microbial genome sequencing. *Genome research*, 15(12), pp.1603-10.

Fredenburgh, J.L., Grizzle, W.D. & Myers, R.B., 2008. Fixation of tissues. *Theory and practice of histological techniques*. Elsevier Health Sciences, p. 725.

Freeman, 2002. *Potential biological control agents for the salmon louse *Lepeophtheirus salmo* (Krøyer 1837)*. University of Sterling.

Freeman, M.A. & Sommerville, C., 2009. *Desmozoon lepeophtherii* n. gen., n. sp., (Microsporidia: Enterocytozoonidae) infecting the salmon louse *Lepeophtheirus salmonis* (Copepoda: Caligidae). *Parasites & vectors*, 2(1), p.58.

Freeman, M.A. & Sommerville, C., 2011. Original observations of *Desmozoon lepeophtherii*, a microsporidian hyperparasite infecting the salmon louse *Lepeophtheirus salmonis*, and its subsequent detection by other researchers. *Parasites & vectors*, 4(1), p.231.

Freeman, M.A., Bell, A.S. & Sommerville, C., 2003. A hyperparasitic microsporidian infecting the salmon louse, *Lepeophtheirus salmonis*: an

rDNA-based molecular phylogenetic study. *Journal of fish diseases*, 26(11-12), pp.667-76.

Freeman, M.A., Kasper, J.M. & Kristmundsson, Á., 2013. *Nucleospora cyclopteri* n. sp., an intranuclear microsporidian infecting wild lumpfish, *Cyclopterus lumpus* L., in Icelandic waters. *Parasites & vectors*, 6, p.49.

Freeman, M.A., Yokoyama, H. & Ogawa, K., 2004. A microsporidian parasite of the genus *Spraguea* in the nervous tissues of the Japanese anglerfish *Lophius litulon*. *Folia parasitologica*, 51(2-3), pp.167-76.

Frixione, E., Ruiz, L., Cerbón, J. & Undeen, A.H., 1997. Germination of *Nosema algerae* (Microspora) spores: conditional inhibition by D2O, ethanol and Hg²⁺ suggests dependence of water influx upon membrane hydration and specific transmembrane pathways. *The Journal of eukaryotic microbiology*, 44(2), pp.109-16.

Frixione, E., Ruiz, L., Santillán, M., de Vargas, L.V., Tejero, J.M. & Undeen, A.H., 1992. Dynamics of Polar Filament Discharge and Sporoplasm Expulsion by Microsporidian Spores. *Cytoskeleton*, 50, pp.38-50.

Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S. & Futcher, B., 2014. Measurement of average decoding rates of the 61 sense codons in vivo. *eLife*, 3.

Geiger, A., Hirtz, C., Bécue, T., Bellard, E., Centeno, D., Gargani, D., Rossignol, M., Cuny, G. & Peltier, J.B., 2010. Exocytosis and protein secretion in *Trypanosoma*. *BioMed central microbiology*, 10(1), p.20.

George, B., Coates, T., McDonald, S., Russ, G., Cherian, S., Nolan, J. & Brealey, J., 2012. Disseminated microsporidiosis with *Encephalitozoon* species in a renal transplant recipient. *Nephrology*, 17, pp.5-8.

Gerlach, S.A., Ekström, D.K. & Eckardt, P.B., 1976. Filter feeding in the hermit crab. *Oecologia*, 24(3), pp.257-264.

German-Retana, S., Candresse, T., Alias, E., Delbos, R.P. & Le Gall, O., 2000. Effects of green fluorescent protein or beta-glucuronidase tagging on the accumulation and pathogenicity of a resistance-breaking Lettuce mosaic

virus isolate in susceptible and resistant lettuce cultivars. *Molecular plant-microbe interactions*, 13(3), pp.316-24.

Germot, A., Philippe, H. & Le Guyader, H., 1997. Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in *Nosema locustae*. *Molecular and biochemical parasitology*, 87(2), pp.159-168.

Ghosh, K., Cappiello, C.D., McBride, S.M., Occi, J.L., Cali, A., Takvorian, P.M., McDonald, T.V. & Weiss, L.M., 2006. Functional characterization of a putative aquaporin from *Encephalitozoon cuniculi*, a microsporidia pathogenic to humans. *International journal for parasitology*, 36(1), pp.57-62.

Gill, E.E. & Fast, N.M., 2006. Assessing the microsporidia-fungi relationship: Combined phylogenetic analysis of eight genes. *Gene*, 375, pp.103-9.

Gill, E.E., Becnel, J.J. & Fast, N.M., 2008. ESTs from the microsporidian *Edhazardia aedis*. *BioMed central genomics*, 9(1), p.296.

Goetz, M., Eichenlaub, S., Pape, G.R. & Hoffmann, R.M., 2001. Chronic diarrhea as a result of intestinal microsporidiosis in a liver transplant recipient. *Transplantation*, 71(2), pp.334-7.

Goldberg, A.V., Molik, S., Tsaousis, A.D., Neumann, K., Kuhnke, G., Delbac, F., Vivares, C.P., Hirt, R.P., Lill, R. & Embley, T.M., 2008. Localization and functionality of microsporidian iron-sulphur cluster assembly proteins. *Nature*, 452(7187), pp.624-8.

Görtz, H.-D., 1986. Endonucleobiosis in Ciliates. *International Review of Cytology*, 102, pp.169–213.

Gouy, M., Guindon, S. & Gascuel, O., 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution*, 27(2), pp.221-4.

Graczyk, T.K., Conn, D.B., Lucy, F., Minchin, D., Tamang, L., Moura, L.N. & DaSilva, A.J., 2004. Human waterborne parasites in zebra mussels (*Dreissena polymorpha*) from the Shannon River drainage area, Ireland. *Parasitology research*, 93(5), pp.385-91.

Gray, M.W., Burger, G. & Lang, B.F., 1999. Mitochondrial Evolution. *Science*,

283(5407), pp.1476-1481.

Gresoviac, S.J., Khattra, J.S., Nadler, S.A., Kent, M.L., Devlin, R.H., Vivares, C.P. & Hedrick, R.P., 2000. Comparison of Small Subunit Ribosomal RNA Gene and Internal Transcribed Spacer Sequences Among Isolates of the Intranuclear Microsporidian *Nucleospora salmonis*. *The Journal of eukaryotic microbiology*, 47(4), pp.379-387.

Grossbard, L. & Schimke, R.T., 1966. Multiple hexokinases of rat tissues. Purification and comparison of soluble forms. *The Journal of biological chemistry*, 241(15), pp.3546-60.

Guerard, A., Rabodonirina, M., Cotte, L., Liguory, O., Piens, M.A., Daoud, S., Picot, S. & Touraine, J.L., 1999. Intestinal microsporidiosis occurring in two renal transplant recipients treated with mycophenolate mofetil. *Transplantation*, 68(5), pp.699-707.

Guillemaud, T., Raymond, M., Tsagkarakou, A., Bernard, C., Rochard, P. & Pasteur, N., 1999. Quantitative variation and selection of esterase gene amplification in *Culex pipiens*. *Heredity*, 83, pp.87-99.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G., 2013. QUASt: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), pp.1072-5.

Gurley, R.R., 1893. On the classification of the Myxosporidia, a group of protozoan parasites infesting fishes. *Bulletin of the United States fish commission*, 11, pp.407-420.

Haag, K.L., James, T.Y., Pombert, J.F., Larsson, R., Schaer, T.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*, (11), pp.1-6.

Haag, K.L., Traunecker, E. & Ebert, D., 2013. Single-nucleotide polymorphisms of two closely related microsporidian parasites suggest a clonal population expansion after the last glaciation. *Molecular ecology*, 22(2), pp.314-26.

Hacker, C., Howell, M., Bhella, D. & Lucocq, J., 2014. Strategies for

maximizing ATP supply in the microsporidian *Encephalitozoon cuniculi*: direct binding of mitochondria to the parasitophorous vacuole and clustering of the mitochondrial porin VDAC. *Cellular microbiology*, 16(4), pp.565-79.

Haferkamp, I., Schmitz-Esser, S., Linka, N., Urbany, C., Collingro, A., Wagner, M., Horn, M. & Neuhaus, H.E., 2004. A candidate NAD⁺ transporter in an intracellular bacterial symbiont related to Chlamydiae. *Nature*, 432(7017), pp.622-625.

Hansen, T. & Schönheit, P., 2003. ATP-dependent glucokinase from the hyperthermophilic bacterium *Thermotoga maritima* represents an extremely thermophilic ROK glucokinase with high substrate specificity. *Federation of European microbiological societies microbiology letters*, 226(2), pp.405-11.

Haro, M., Henriques-Gil, N., Fenoy, S., Izquierdo, F., Alonso, F. & Del Aguila, C., 2006. Detection and genotyping of *Enterocytozoon bieneusi* in pigeons. *Journal of eukaryotic microbiology*, 53, pp.S58-60.

Haro, M., Izquierdo, F., Henriques-Gil, N., Andrés, I., Alonso, F., Fenoy, S. & Del Aguila, C., 2005. First detection and genotyping of human-associated microsporidia in pigeons from urban parks. *Applied and environmental microbiology*, 71(6), pp.3153-7.

Hedrick R.P., Groff J.M., McDowell T.S., Willis M. & Cox W.T., 1990. Hematopoietic intranuclear microsporidian infections with features of leukemia in chinook salmon *Oncorhynchus tshawytscha*. *Diseases of aquatic organisms*, 8, pp.189-197.

Hedrick, R.P., Groff, J.M. & Baxa, D. V., 1991. Experimental infections with *Nucleospora salmonis* n. g. n. sp.: An intranuclear microsporidium from chinook salmon (*Oncorhynchus tshawytscha*). *American Fisheries Society, Fish Health Section Newsletter*, 19, p.5.

Heinz, E., Hacker, C., Dean, P., Mifsud, J., Goldberg, A.V., Williams, T.A., Nakjang, S., Gregory, A., Hirt, R.P., Lucocq, J.M. & Kunji, E.R., 2014. Plasma membrane-located purine nucleotide transport proteins are key components for host exploitation by microsporidian intracellular parasites. *Public library of science pathogens*, 10(12).

Heinz, E., Williams, T.A., Nakjang, S., Noël, C.J., Swan, D.C., Goldberg, A.V., Harris, S.R., Weinmaier, T., Markert, S., Becher, D., & Bernhardt, J., 2012. The genome of the obligate intracellular parasite *Trachipleistophora hominis*: new insights into microsporidian genome dynamics and reductive evolution. *Public library of science pathogens*, 8(10).

Henriques-Gil, N., Haro, M., Izquierdo, F., Fenoy, S. & del Águila, C., 2010. Phylogenetic approach to the variability of the microsporidian *Enterocytozoon bieneusi* and its implications for inter- and intrahost transmission. *Applied and environmental microbiology*, 76(10), pp.3333-42.

Henze, K., Horner, D.S., Suguri, S., Moore, D.V., Sánchez, L.B., Müller, M. & Embley, T.M., 2001. Unique phylogenetic relationships of glucokinase and glucosephosphate isomerase of the amitochondriate eukaryotes *Giardia intestinalis*, *Spironucleus barkhanus* and *Trichomonas vaginalis*. *Gene*, 281(1-2), pp.123-131.

Hershberg, R. & Petrov, D.A., 2008. Selection on codon bias. *Annual review of genetics*, 42, pp.287-99.

Higgins, M.J., Kent, M.L., Moran, J.D.W., Weiss, L.M. & Dawe, S.C., 1998. Efficacy of the fumagillin analog TNP-470 for *Nucleospora salmonis* and *Loma salmonae* infections in chinook salmon *Oncorhynchus tshawytscha*. *Diseases of aquatic organisms*, 34(1), pp.45-9.

Hiller, N.L., Bhattacharjee, S., van Ooij, C., Liolios, K., Harrison, T., Lopez-Estrano, C. & Haldar, K., 2004. A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science*, 306(5703), pp.1934-7.

Hinkle, G., Leipe, D.D., Nerad, T.A. & Sogin, M.L., 1994. The unusually long small subunit ribosomal RNA of *Phreatamoeba balamuthi*. *Nucleic acids research*, 22(3), pp.465-9.

Hinkle, G., Morrison, H.G. & Sogin, M.L., 1997. Genes coding for reverse transcriptase, DNA-directed RNA polymerase, and chitin synthase from the microsporidian *Spraguea lophii*. *The Biological bulletin*, 193(2), pp.250-1.

Hirt, R.P., Healy, B., Vossbrinck, C.R., Canning, E.U. & Embley, T.M., 1997. A mitochondrial Hsp70 orthologue in *Vairimorpha necatrix*: molecular

evidence that microsporidia once contained mitochondria. *Current biology*, 7(12), pp.995-998.

Hirt, R.P., Logsdon, J.M., Healy, B., Dorey, M.W., Doolittle, W.F. & Embley, T.M., 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proceedings of the national academy of sciences of the United States of America*, 96(2), pp.580-585.

Hofstetter, V., Miadlikowska, J., Kauff, F. & Lutzoni, F., 2007. Phylogenetic comparison of protein-coding versus ribosomal RNA-coding sequence data: a case study of the *Lecanoromycetes* (Ascomycota). *Molecular phylogenetics and evolution*, 44(1), pp.412-26.

Hohmann, S., Winderickx, J., de Winde, J.H., Valckx, D., Cobbaert, P., Luyten, K., de Meirman, C., Ramos, J. & Thevelein, J.M., 1999. Novel alleles of yeast hexokinase PII with distinct effects on catalytic activity and catabolite repression of SUC2. *Microbiology*, pp.703-14.

Hollister, W.S., Canning, E.U., Weidner, E., Field, A.S., Kench, J. & Marriott, D.J., 1996. Development and ultrastructure of *Trachipleistophora hominis* n.g., n.sp. after in vitro isolation from an AIDS patient and inoculation into athymic mice. *Parasitology*, 112(1), pp.143-54.

Hori, M., Fujii, K. & Fujishama, M., 2008. Micronucleus-Specific Bacterium *Holospora elegans* Irreversibly Enhances Stress Gene Expression of the Host *Paramecium caudatum*. *Journal of Eukaryotic Microbiology*, 55(6), pp.515–21.

Horn, H., Keller, A., Hildebrandt, U., Kämpfer, P., Riederer, M. & Hentschel, U., 2016. Draft genome of the *Arabidopsis thaliana* phyllosphere bacterium, *Williamsia* sp. ARP1. *Standards in genomic sciences*, 11(1), p.8.

Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J. & Nakai, K., 2007. WoLF PSORT: protein localization predictor. *Nucleic acids research*, 35, pp.W585-7.

Horwitz, M.A., 1983. Formation of a novel phagosome by the Legionnaires' disease bacterium (*Legionella pneumophila*) in human monocytes. *The Journal of experimental medicine*, 158(4), pp.1319-31.

- Hrdý, I., Mertens, E. & Nohýnková, E., 1993. Giardia intestinalis: detection and characterization of a pyruvate phosphate dikinase. *Experimental parasitology*, 76(4), pp.438-41.
- Huang, J. & Brumell, J.H., 2014. Bacteria–autophagy interplay: a battle for survival. *Nature Reviews Microbiology*, 12(2), pp.101–114.
- Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F. & Wheeler, T.J., 2015. The Dfam database of repetitive DNA families. *Nucleic acids research*, 44(D1), pp.D81-89.
- Husemann, P. & Stoye, J., 2010. r2cat: synteny plots and comparative assembly. *Bioinformatics*, 26(4), pp.570-1.
- Ihaka, R. & Gentleman, R., 2012. R: A Language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3), pp.299-314
- Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology and evolution*, 2(1), pp.13-34.
- Imamura, H., Nhat, K.P.H., Togawa, H., Saito, K., Iino, R., Kato-Yamada, Y., Nagai, T. & Noji, H., 2009. Visualization of ATP levels inside single living cells with fluorescence resonance energy transfer-based genetically encoded indicators. *Proceedings of the national academy of sciences of the United States of America*, 106(37), pp.15651-6.
- Imelfort, M., 2009. Sequence comparison tools. In D. Edwards, D. Hansen, & J. Stajich, eds. *Bioinformatics: Tools and Applications*. London: Springer, pp.13-37.
- Inagaki, Y., Susko, E., Fast, N.M. & Roger, A.J., 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and Archaeobacteria in EF-1 α phylogenies. *Molecular biology and evolution*, 21(7), pp.1340-1349.
- Ingle, R.W., 1980. *British Crabs*, London: Oxford University Press and British Natural History Museum.
- Ironside, J.E., 2013. Diversity and recombination of dispersed ribosomal DNA and protein coding genes in microsporidia. *Public library of science one*, 8(2).

Ishihara, R. & Hayashi, Y., 1968. Some properties of ribosomes from the sporoplasm of *Nosema bombycis*. *Journal of invertebrate pathology*, 11(3), pp.377-385.

Ishikawa, K., Ogawa, T., Hirose, E., Nakayama, Y., Harada, K., Fukusaki, E., Yoshimura, K. & Shigeoka, S., 2009. Modulation of the poly(ADP-ribose)ation reaction via the Arabidopsis ADP-ribose/NADH pyrophosphohydrolase, AtNUDX7, is involved in the response to oxidative stress. *Plant physiology*, 151(2), pp.741-754.

Ismail, T.M., Hart, C.A. & McLennan, A.G., 2003. Regulation of dinucleoside polyphosphate pools by the YgdP and ApaH hydrolases is essential for the ability of *Salmonella enterica* serovar typhimurium to invade cultured mammalian cells. *The Journal of biological chemistry*, 278(35), pp.32602-7.

James, G.A., Korber, D.R., Caldwell, D.E. & Costerton, J.W., 1995. Digital image analysis of growth and starvation responses of a surface-colonizing *Acinetobacter* sp. *J. Bacteriol.*, 177(4), pp.907-915.

James, L.C. & Tawfik, D.S., 2003. Conformational diversity and protein evolution: a 60-year-old hypothesis revisited. *Trends in biochemical sciences*, 28(7), pp.361-8.

James, T.Y., Kauff, F., Schoch, C.L., Matheny, P.B., Hofstetter, V., Cox, C.J., Celio, G., Gueidan, C., Fraker, E., Miadlikowska, J. & Lumbsch, H.T., 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*, 443(7113), pp.818-22.

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*, 23(16), pp.1548-53.

Janssens, V. & Goris, J., 2001. Protein phosphatase 2A: a highly regulated family of serine/threonine phosphatases implicated in cell growth and signalling. *The biochemical journal*, 353, pp.417-39.

Jeong, H., Lee, D.H., Ryu, C.M. & Park, S.H., 2015. Toward complete bacterial genome sequencing through the combined use of multiple next-generation sequencing platforms. *Journal of microbiology and biotechnology*,

26(1), pp.207-212.

Johnson, M.A., Becnel, J.J. & Undeen, A.H., 1997. A new sporulation sequence in *Edhazardia aedis* (Microsporidia: Culicosporidae), a parasite of the mosquito *Aedes aegypti* (Diptera: Culicidae). *Journal of invertebrate pathology*, 70(1), pp.69-75.

Jones, M.B., Rosenberg, J.N., Betenbaugh, M.J. & Krag, S.S., 2009. Structure and synthesis of polyisoprenoids used in N-glycosylation across the three domains of life. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1790(6), pp.485-494.

Jones, M.D., Forn, I., Gadelha, C., Egan, M.J., Bass, D., Massana, R. & Richards, T.A., 2011. Discovery of novel intermediate forms redefines the fungal tree of life. *Nature*, 474(7350), pp.200-3.

Jones, M.D., Richards, T.A., Hawksworth, D.L. & Bass, D., 2011. Validation and justification of the phylum name Cryptomycota phyl. nov. *IMA fungus*, 2(2), pp.173-5.

Jovanovic, A., Alekseev, A.E. & Terzic, A., 1997. Intracellular diadenosine polyphosphates. *Biochemical pharmacology*, 54(2), pp.219-225.

Kamaishi, T., Hashimoto, T., Nakamura, Y., Masuda, Y., Nakamura, F., Okamoto, K.I., Shimizu, M. & Hasegawa, M., 1996. Complete nucleotide sequences of the genes encoding translation elongation factors 1 alpha and 2 from a microsporidian parasite, *Glugea plecoglossi*: implications for the deepest branching of eukaryotes. *Journal of biochemistry*, 120(6), pp.1095-103.

Kamaishi, T., Hashimoto, T., Nakamura, Y., Nakamura, F., Murata, S., Okada, N., Okamoto, K.I., Shimizu, M. & Hasegawa, M., 1996. Protein phylogeny of translation elongation factor EF-1 alpha suggests microsporidians are extremely ancient eukaryotes. *Journal of molecular evolution*, 42(2), pp.257-63.

Karpov, S.A., Mikhailov, K.V., Mirzaeva, G.S., Mirabdullaev, I.M., Mamkaeva, K.A., Titova, N.N. & Aleoshin, V.V., 2013. Obligately phagotrophic aphelids turned out to branch with the earliest-diverging fungi. *Protist*, 164(2), pp.195-

205.

Katinka, M.D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretailade, E., Brottier, P., Wincker, P. & Delbac, F., 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*, 414.

Kato, N., Mueller, C.R., Fuchs, J.F., Wessely, V., Lan, Q. & Christensen, B.M., 2006. Regulatory mechanisms of chitin biosynthesis and roles of chitin in peritrophic matrix formation in the midgut of adult *Aedes aegypti*. *Insect biochemistry and molecular biology*, 36(1), pp.1-9.

Katzen, H.M. & Schimke, R.T., 1965. Multiple forms of hexokinase in the rat: tissue distribution, age dependency, and properties. *Proceedings of the national academy of sciences of the United States of America*, 54(4), pp.1218-25.

Kearney, J., 2010. Food consumption trends and drivers. *Philosophical transactions of the Royal Society of London*, 365(1554), pp.2793-807.

Keating, L.A., Wheeler, P.R., Mansoor, H., Inwald, J.K., Dale, J., Hewinson, R.G. & Gordon, S.V., 2005. The pyruvate requirement of some members of the *Mycobacterium tuberculosis* complex is due to an inactive pyruvate kinase: implications for *in vivo* growth. *Molecular microbiology*, 56(1), pp.163-74.

Keeling, P.J. & Corradi, N., 2011. Shrink it or lose it: balancing loss of function with shrinking genomes in the microsporidia. *Virulence*, 2(1), pp.67-70.

Keeling, P.J. & Doolittle, W.F., 1996. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. *Molecular biology and evolution*, 13(10), pp.1297-305.

Keeling, P.J. & Fast, N.M., 2002. Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annual review of microbiology*, 56, pp.93-116.

Keeling, P.J. & McFadden, G.I., 1998. Origins of microsporidia. *Trends in microbiology*, 6(1), pp.19-23.

Keeling, P.J., 2003. Congruent evidence from α -tubulin and β -tubulin gene phylogenies for a zygomycete origin of microsporidia. *Fungal genetics and biology*, 38(3), pp.298-309.

Keeling, P.J., 2014. Phylogenetic place of Microsporidia in the tree of eukaryotes. In L. M. Weiss & J. J. Becnel, eds. *Microsporidia: Pathogens of opportunity*. John Wiley and Sons, inc., pp. 195-202.

Keeling, P.J., Burger, G., Durnford, D.G., Lang, B.F., Lee, R.W., Pearlman, R.E., Roger, A.J. & Gray, M.W., 2005. The tree of eukaryotes. *Trends in ecology & evolution*, 20(12), pp.670-6.

Keeling, P.J., Corradi, N., Morrison, H.G., Haag, K.L., Ebert, D., Weiss, L.M., Akiyoshi, D.E. & Tzipori, S., 2010. The reduced genome of the parasitic microsporidian *Enterocytozoon bieneusi* lacks genes for core carbon metabolism. *Genome biology and evolution*, 2, pp.304-9.

Keeling, P.J., Luker, M.A. & Palmer, J.D., 2000. Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. *Molecular biology and evolution*, 17(1), pp.23-31.

Kelkar, R., Sastry, P.S.R.K., Kulkarni, S.S., Saikia, T.K., Parikh, P.M. & Advani, S.H., 1997. Pulmonary microsporidial infection in a patient with CML undergoing allogeneic marrow transplant. *Bone Marrow Transplantation*, 19(2), pp.179-182.

Kelkar, Y.D. & Ochman, H., 2012. Causes and consequences of genome expansion in fungi. *Genome biology and evolution*, 4(1), pp.13-23.

Kent, M.L., Hervio, D.M., Docker, M.F. & Devlin, R.H., 1996. Taxonomy studies and diagnostic tests for myxosporean and microsporidian pathogens of salmonid fishes utilising ribosomal DNA sequence. *The Journal of eukaryotic microbiology*, 43(5), pp.98S-99S.

Kent, M.L., Shaw, R.W. & Sanders, J.L., 2014. Microsporidia in Fish. In L. M. Weiss & J. J. Becnel, eds. *Microsporidia: Pathogens of Opportunity*. Pondicherry: John Wiley & Sons, Inc., pp. 493-520.

Khatri, B., Fielder, M., Jones, G., Newell, W., Abu-Oun, M. & Wheeler, P.R.,

2013. High Throughput Phenotypic Analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* strains' metabolism using biolog phenotype microarrays. *Public library of science one*, 8(1).

Khattra, J.S., Gresoviac, S.J., Kent, M.L., Myers, M.S., Hedrick, R.P. & Devlin, R.H., 2000. Molecular detection and phylogenetic placement of a microsporidian from English sole (*Pleuronectes vetulus*) affected by X-cell pseudotumors. *The Journal of parasitology*, 86(4), pp.867-71.

Kim, J.-W. & Dang, C. V, 2005. Multifaceted roles of glycolytic enzymes. *Trends in biochemical sciences*, 30(3), pp.142-50.

Kitano, H., 2007. Towards a theory of biological robustness. *Molecular systems biology*, 3(1), p.137.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I. & Koonin, E.V., 2002. Selection in the evolution of gene duplications. *Genome biology*, 3(2).

Koren, S., Treangen, T.J., Hill, C.M., Pop, M. & Phillippy, A.M., 2014. Automated ensemble assembly and validation of microbial genomes. *BioMed central bioinformatics*, 15(1), p.126.

Krebs, H.A., 1970. The history of the tricarboxylic acid cycle. *Perspectives in biology and medicine*, 14(1), pp.154–70.

Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*, 305(3), pp.567-80.

Kubitschek, H.E., 1990. Cell volume increase in *Escherichia coli* after shifts to richer media. *Journal of bacteriology*, 172(1), pp.94-101.

Kudo, R.R. & Daniels, E.W., 1963. An electron microscope study of the spore of the microsporidian *Thelohania californica*. *Journal of protozoology*, 10, pp.112-20.

Kudo, R.R., 1920. On the structure of some microsporidian spores. *Journal of parasitology*, 6(4), pp.178-182.

Kudo, R.R., 1921. On the nature of structures characteristic of cnidosporidian

spores. *Transactions of the American microscopical society*, 40(2), pp.59-74.

Kudo, R.R., 1924. Taxonomy. In *A biologic and taxonomic study of the Microsporidia*. Illinois, pp. 61-199.

Kudo, R.R., 1930. Studies on microsporidia parasitic in mosquitoes. VIII. On a microsporidian, *Nosema aedis* nov. spec., parasitic in a larva of *Aedes aegypti* of Puerto Rico. *Archiv fur Protistenkunde*, 69, pp.23-28.

Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. & Blaxter, M., 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in genetics*, 4, p.237.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S.L., 2004. Versatile and open software for comparing large genomes. *Genome biology*, 5(2), p.R12.

Labbé, A., 1899. Sporozoa. *Das Tierreich*. Berlin: Friedlander u Sohn, p. 180.

Lai, Y.T., Chang, Y.Y., Hu, L., Yang, Y., Chao, A., Du, Z.Y., Tanner, J.A., Chye, M.L., Qian, C., Ng, K.M. & Li, H., 2015. Rapid labeling of intracellular His-tagged proteins in living cells. *Proceedings of the national academy of sciences of the United States of America*, 112(10), pp.2948-53.

Larsson, J.I., 1989. Light and electron microscope studies on *Jirovecia involuta* sp. nov. (Microspora, Bacillidiidae), a new microsporidian parasite of oligochaetes in Sweden. *European journal of protistology*, 25(2), pp.172-81.

Larsson, J.I.R., 1986. Ultrastructure, function, and classification of Microsporidia. *Progress in protistology*, 1, pp.325-390.

Larsson, J.I.R., 1994. Characteristics of the genus *Bacillidium* Janda, 1928 (Microspora, Mrazekiidae)-Reinvestigation of the type species *B. criodrili* and improved diagnosis of the genus. *European journal of protistology*, 30(1), pp.85-96.

Larsson, R., 1980. Insect pathological investigations on Swedish *Thysanura*. II. A new microsporidian parasite of *Petrobius brevistylis* (Microcoryphia, Machilidae); description of the species and creation of two new genera and a

new family. *Protistologica*.

Lavner, Y. & Kotlar, D., 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene*, 345(1), pp.127-38.

Law, R.J., Alaei, M., Allchin, C.R., Boon, J.P., Lebeuf, M., Lepom, P. & Stern, G.A., 2003. Levels and trends of polybrominated diphenylethers and other brominated flame retardants in wildlife. *Environment international*, 29(6), pp.757-70.

Lawton, P., 1989. Predatory interaction between the brachyuran crab *Cancer pagurus* and decapod crustacean prey. *Marine ecology progress series*, 52(2), pp.169-179.

Lee, J.S. & Pfeifer, D.K., 1977. Microbiological characteristics of Pacific shrimp (*Pandalus jordani*). *Applied and environmental microbiology*, 33(4), pp.853-9.

Lee, S.C., Corradi, N., Byrnes, E.J., Torres-Martinez, S., Dietrich, F.S., Keeling, P.J. & Heitman, J., 2008. Microsporidia evolved from ancestral sexual fungi. *Current biology*, 18(21), pp.1675-9.

Lee, S.C., Weiss, L.M. & Heitman, J., 2009. Generation of genetic diversity in microsporidia via sexual reproduction and horizontal gene transfer. *Communicative & integrative biology*, 2(5), pp.414-7.

Lee, Y.N., Nechushtan, H., Figov, N. & Razin, E., 2004. The function of lysyl-tRNA synthetase and Ap4A as signaling regulators of MITF activity in ϵ -activated mast cells. *Immunity*, 20(2), pp.145-51.

Leelayoova, S., Subrungruang, I., Rangsin, R., Chavalitshewinkoon-Petmitr, P., Worapong, J., Naaglor, T. & Mungthin, M., 2005. Transmission of *Enterocytozoon bieneusi* genotype a in a Thai orphanage. *The American journal of tropical medicine and hygiene*, 73(1), pp.104-7.

Letcher, P.M., Lopez, S., Schmieder, R., Lee, P.A., Behnke, C., Powell, M.J. & McBride, R.C., 2013. Characterization of *Amoebophilidium protococcarum*, an algal parasite new to the Cryptomycota isolated from an outdoor algal pond used for the production of biofuel. *Public library of science*

one, 8(2).

Li, L., Stoeckert, C.J. & Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9), pp.2178-89.

Li, Y.L., Weng, J.C., Hsiao, C.C., Chou, M.T., Tseng, C.W. & Hung, J.H., 2015. PEAT: an intelligent and efficient paired-end sequencing adapter trimming algorithm. *BioMed central bioinformatics*, 16(1), p.1

Liu, H., Pan, G., Song, S., Xu, J., Li, T., Deng, Y. & Zhou, Z., 2008. Multiple rDNA units distributed on all chromosomes of *Nosema bombycis*. *Journal of invertebrate pathology*, 99(2), pp.235-8.

Liu, T.P., 1972. Ultrastructural changes in the nuclear envelope of larval fat body cells of *Simulium vittatum* (Diptera) induced by microsporidian infection of *Thelohania bracteata*. *Tissue & cell*, 4(3), pp.493-501.

Lom, J. & Dykoá, I., 2002. Ultrastructure of *Nucleospora secunda* n. sp. (Microsporidia), parasite of enterocytes of *Nothobranchius rubripinnis*. *European Journal of Protistology*, 38(1), pp.19-27.

Longshaw, M., Feist, S.W. & Bateman, K.S., 2012. Parasites and pathogens of the endosymbiotic pea crab (*Pinnotheres pisum*) from blue mussels (*Mytilus edulis*) in England. *Journal of invertebrate pathology*, 109(2), pp.235-42.

López-Vélez, R., Turrientes, M.C., Garrón, C., Montilla, P., Navajas, R., Fenoy, S. & del Aguila, C., 1999. Microsporidiosis in travelers with diarrhea from the tropics. *Journal of travel medicine*, 6(4), pp.223-7.

Lores, B., del Aguila, C. & Arias, C., 2002. *Enterocytozoon bieneusi* (microsporidia) in faecal samples from domestic animals from Galicia, Spain. *Memórias do instituto Oswaldo Cruz*, 97(7), pp.941-5.

Loubès, C. & Akbarieh, M., 1978. Étude ultrastructurale de la Microsporidie *Baculea daphniae* n. g, n. sp., parasite de l'épithélium intestinal de *Daphnia pulex* Leydig, 1860 (Crustacé, Cladocère). *Protistologica*, 14, pp.23-38.

Lowe, T.M. & Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids*

research, 25(5), pp.955-64.

Lücking, R., Huhndorf, S., Pfister, D.H., Plata, E.R. & Lumbsch, H.T., 2009. Fungi evolved right on track. *Mycologia*, 101(6), pp.810-22.

Lynch, M., O'Hely, M., Walsh, B. & Force, A., 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics*, 159(4), pp.1789-804.

Machanick, P. & Bailey, T.L., 2011. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27(12), pp.1696-7.

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., Tallon, L.J. & Salzberg, S.L., 2013. GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, 29(14), pp.1718-25.

Mahen, R., Koch, B., Wachsmuth, M., Politi, A.Z., Perez-Gonzalez, A., Mergenthaler, J., Cai, Y. & Ellenberg, J., 2014. Comparative assessment of fluorescent transgene methods for quantitative imaging in human cells. *Molecular biology of the cell*, 25(22), pp.3610-8.

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C. & Bryant, S.H., 2015. CDD: NCBI's conserved domain database. *Nucleic acids research*, 43, pp.D222-6.

Mardis, E.R., 2006. Anticipating the 1,000 dollar genome. *Genome biology*, 7(7), p.112.

Mardis, E.R., 2008. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*, 9, pp.387-402.

Mardis, E.R., 2009. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. *Genome medicine*, 1(4), p.40.

Margulis, L., 1970. *Origin of eukaryotic cells: evidence and research implications for a theory of the origin and evolution of microbial, plant, and animal cells on the Precambrian earth*. New Haven: Yale University Press.

Margulis, L., 1981. *Symbiosis in cell evolution: Life and its environment on the early earth*. San Francisco: W. H. Freeman and Co.

Marques-Bonet, T., Girirajan, S. & Eichler, E.E., 2009. The origins and impact of primate segmental duplications. *Trends in genetics*, 25(10), pp.443-54.

Martín, F., Pintor, J., Rovira, J.M., Ripoll, C., Miras-Portugal, M.T. & Soria, B., 1998. Intracellular diadenosine polyphosphates: a novel second messenger in stimulus-secretion coupling. *Federation of American Societies for Experimental Biology*, 12(14), pp.1499-506.

Mathews, C.K., 2015. Deoxyribonucleotide metabolism, mutagenesis and cancer. *Nature reviews cancer*, 15(9), pp.528-39.

Mathis, A., Weber, R. & Deplazes, P., 2005. Zoonotic potential of the microsporidia. *Clinical microbiology reviews*, 18(3), pp.423-45.

Matos, O., Lobo, M.L. & Xiao, L., 2012. Epidemiology of *Enterocytozoon bieneusi* Infection in humans. *Journal of parasitology research*, 2012, p.981424.

Matsumoto, A., Bessho, H., Uehira, K. & Suda, T., 1991. Morphological studies of the association of mitochondria with chlamydial inclusions and the fusion of chlamydial inclusions. *Journal of electron microscopy*, 40(5), pp.356-63.

Matthew. E., Ellis, G., Alistair. M., Pilgrim, S., Reade, S., Williamson, K. & Wintz, P., 2015. Landings. *UK Sea fisheries statistics 2014*. London: National Statistics.

Mayordomo, I. & Sanz, P., 2001. Hexokinase PII: structural analysis and glucose signalling in the yeast *Saccharomyces cerevisiae*. *Yeast*, 18(10), pp.923-30.

Mendonça, A.G., Alves, R.J. & Pereira-Leal, J.B., 2011. Loss of genetic redundancy in reductive genome evolution. *Public library of science computational biology*, 7(2).

Merhej, V., Royer-Carenzi, M., Pontarotti, P. & Raoult, D., 2009. Massive comparative genomic analysis reveals convergent evolution of specialized

bacteria. *Biology direct*, 4, p.1.

Miller, D.S. & Horowitz, S.B., 1986. Intracellular compartmentalization of adenosine triphosphate. *The Journal of biological chemistry*, 261(30), pp.13911–5.

Miller, J.J., Delwiche, C.F. & Coats, D.W., 2012. Ultrastructure of *Amoebophrya* sp. and its Changes during the Course of Infection. *Protist*, 163(5), pp.720–45.

Millward, D.J., Garlick, P.J., Stewart, R.J., Nnanyelugo, D.O. & Waterlow, J.C., 1975. Skeletal-muscle growth and protein turnover. *Biochemical Journal*, 150(2), pp.235-243.

Mirebrahim, H., Close, T.J. & Lonardi, S., 2015. *De novo* meta-assembly of ultra-deep sequencing data. *Bioinformatics*, 31(12), pp.9-16.

Modin, J.C., 1981. *Microsporidium rhabdophilia* n. sp. from rodlet cells of salmonid fishes. *Journal of fish diseases*, 4(3), pp.203-211.

Monaghan, S.R., Rumney, R.L., Vo, N.T., Bols, N.C. & Lee, L.E., 2011. *In vitro* growth of microsporidia *Anncaliia algerae* in cell lines from warm water fish. *In vitro cellular & developmental biology-Animal*, 47(2), pp.104-13.

Mori, H., Mahittikorn, A., Watthanakulpanich, D., Komalamisra, C. & Sukthana, Y., 2013. Zoonotic potential of *Enterocytozoon bieneusi* among children in rural communities in Thailand. *Parasite*, 20, p.14.

Morin, L. & Mignot, J.P., 1995. "Are Archamoebae true Archezoa? The phylogenetic position of *Pelomyxa* sp., as inferred from large subunit ribosomal RNA sequencing." *European journal of protistology*, 31(402).

Morris, S., Allchin, C.R., Zegers, B.N., Haftka, J.J., Boon, J.P., Belpaire, C., Leonards, P.E., Van Leeuwen, S.P. & de Boer, J., 2004. Distribution and fate of HBCD and TBBPA brominated flame retardants in North Sea estuaries and aquatic food webs. *Environmental science & technology*, 38(21), pp.5497-5504.

Morrison, J.K., MacConnell, E., Chapman, P.F. & Westgard, R.L., 1990. A microsporidium-induced lymphoblastosis in chinook salmon *Oncorhynchus*

- tshawytscha*. *Diseases of aquatic organisms*, 8(2), pp.99-104.
- Mount, D.W., 2007. Using the Basic Local Alignment Search Tool (BLAST). *Cold spring harbour protocols*, 2007, p.pdb.top17.
- Müller, A., Bialek, R., Kämper, A., Fätkenheuer, G., Salzberger, B. & Franzen, C., 2001. Detection of microsporidia in travelers with diarrhea. *Journal of clinical microbiology*, 39(4), pp.1630-2.
- Müller, A., Trammer, T., Chioralia, G., Seitz, H.M., Diehl, V. & Franzen, C., 2000. Ribosomal RNA of *Nosema algerae* and phylogenetic relationship to other microsporidia. *Parasitology research*, 86(1), pp.18-23.
- Mullins, J.E., Powell, M., Speare, D.J. & Cawthorn, R.J., 1994. An intranuclear microsporidian in lumpfish *Cyclopterus lumpus*. *Diseases of aquatic organisms*, 20(1), p.7.
- Muñoz-López, M. & García-Pérez, J.L., 2010. DNA transposons: nature and applications in genomics. *Current genomics*, 11(2), pp.115-28.
- Myers, E.W., 1995. Toward simplifying and accurately formulating fragment assembly. *Journal of computational biology*, 2(2), pp.275-90.
- Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A. & Anson, E.L., 2000. A whole-genome assembly of *Drosophila*. *Science*, 287(5461), pp.2196-204.
- Nagarajan, N. & Pop, M., 2009. Parametric complexity of sequence assembly: theory and applications to next generation sequencing. *Journal of computational biology*, 16(7), pp.897-908.
- Nagarajan, N. & Pop, M., 2013. Sequence assembly demystified. *Nature reviews genetics*, 14(3), pp.157-67.
- 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), pp.2285-303.

Naora, H., Naora, H., Izawa, M., Allfrey, V.G. & Mirski, A.E., 1962. Some observations on differences in composition between the nucleus and cytoplasm of the frog oocyte. *Proceedings of the National Academy of Sciences of the United States of America*, 48(5), pp.853–9.

Nascimento, R., Gouran, H., Chakraborty, S., Gillespie, H.W., Almeida-Souza, H.O., Tu, A., Rao, B.J., Feldstein, P.A., Bruening, G., Goulart, L.R. & Dandekar, A.M., 2016. The Type II Secreted Lipase/Esterase LesA is a Key Virulence Factor Required for *Xylella fastidiosa* Pathogenesis in Grapevines. *Scientific reports*, 6, p.18598.

Néron, B., Ménager, H., Maufrais, C., Joly, N., Maupetit, J., Letort, S., Carrere, S., Tuffery, P. & Letondal, C., 2009. Mobylye: a new full web bioinformatics framework. *Bioinformatics*, 25(22), pp.3005-11.

Nilsson, A., Olsson, T., Ulfstedt, M., Thelander, M. & Ronne, H., 2011. Two novel types of hexokinases in the moss *Physcomitrella patens*. *BioMed central Plant Biology*, 11(1), p.32.

Noji, H. & Yoshida, M., 2001. The Rotary Machine in the Cell, ATP Synthase. *Journal of Biological Chemistry*, 276(3), pp.1665–8.

Nowak, M.A., Boerlijst, M.C., Cooke, J. & Smith, J.M., 1997. Evolution of genetic redundancy. *Nature*, 388(6638), pp.167-71.

Nylund, S., Nylund, A., Watanabe, K., Arnesen, C.E. & Karlsbakk, E., 2010. *Paranucleospora theridion* n. gen., n. sp. (Microsporidia, Enterocytozoonidae) with a life cycle in the salmon louse (*Lepeophtheirus salmonis*, Copepoda) and Atlantic salmon (*Salmo salar*). *The Journal of eukaryotic microbiology*, 57(2), pp.95-114.

Oakley, T.H., Ostman, B. & Wilson, A.C. V., 2006. Repression and loss of gene expression outpaces activation and gain in recently duplicated fly genes. *Proceedings of the national academy of sciences*, 103(31), pp.11637-41.

Ogawa, T., Muramoto, K., Takada, R., Nakagawa, S., Shigeoka, S. & Yoshimura, K., 2016. Modulation of NADH levels by *Arabidopsis* NUDIX hydrolases, AtNUDX6 and 7, and the respective proteins themselves play distinct roles in the regulation of various cellular responses involved in

biotic/abiotic stresses. *Plant & cell physiology*, 57(6), pp.1295-308.

Ondov, B.D., Bergman, N.H. & Phillippy, A.M., 2011. Interactive metagenomic visualization in a web browser. *BioMed central bioinformatics*, 12(1), p.1.

Opperdoes, F.R., De Jonckheere, J.F. & Tielens, A.G.M., 2011. *Naegleria gruberi* metabolism. *International journal for parasitology*, 41(9), pp.915-24.

Orenstein, J.M., Russo, P., Didier, E.S., Bowers, C., Bunin, N. & Teachey, D.T., 2005. Fatal pulmonary microsporidiosis due to *Encephalitozoon cuniculi* following allogeneic bone marrow transplantation for acute myelogenous leukemia. *Ultrastructural pathology*, 29(3-4), pp.269-76.

Overstreet, R.M. & Weidner, E., 1974. Differentiation of microsporidian spore-tails in *Inodosporus spraguei* gen. et sp. n. *Zeitschrift für Parasitenkunde*, 44(3), pp.169-86.

Palenzuela, O., Redondo, M.J., Cali, A., Takvorian, P.M., Alonso-Naveiro, M., Alvarez-Pellitero, P. & Sitjà-Bobadilla, A., 2014. A new intranuclear microsporidium, *Enterospora nucleophila* n. sp., causing an emaciative syndrome in a piscine host (*Sparus aurata*), prompts the redescription of the family Enterocytozoonidae. *International journal for parasitology*, 44(3-4), pp.189-203.

Pan, G., Xu, J., Li, T., Xia, Q., Liu, S.L., Zhang, G., Li, S., Li, C., Liu, H., Yang, L. & Liu, T., 2013. Comparative genomics of parasitic silkworm microsporidia reveal an association between genome expansion and host adaptation. *BioMed central genomics*, 14(1), p.186.

Papp, B., Pál, C. & Hurst, L.D., 2004. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, 429(6992), pp.661-4.

Parisot, N., Pelin, A., Gasc, C., Polonais, V., Belkorchia, A., Panek, J., El Alaoui, H., Biron, D.G., Brassat, É., Vauray, C. & Peyret, P., 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), pp.2289-300.

- Park, J.Y., Kim, S., Kim, S.M., Cha, S.H., Lim, S.K. & Kim, J., 2011. Complete genome sequence of multidrug-resistant *Acinetobacter baumannii* strain 1656-2, which forms sturdy biofilm. *Journal of bacteriology*, 193(22), pp.6393-4.
- Patterson, J.H., Waller, R.F., Jeevarajah, D., Billman-Jacobe, H. & McCONVILLE, M.J., 2003. Mannose metabolism is required for mycobacterial growth. *Biochemical journal*, 372, pp.77-86.
- Pavan, C., Perondini, A.L.P. & Picard, T., 1969. Changes in chromosomes and in development of cells of *Sciara ocellaris* induced by microsporidian infections. *Chromosoma*, 28(3).
- Pérez, C., 1905. Sur une *Glugea* nouvelle parasite de *Balanus amaryllis*. *Comptes Rendus des Seances de la Societe de Biologie*, 58, pp.150-151.
- Petersen, A., Andersen, J.S., Kaewmak, T., Somsiri, T. & Dalsgaard, A., 2002. Impact of integrated fish farming on antimicrobial resistance in a pond environment. *Applied and Environmental Microbiology*, 68(12), pp.6036-42.
- Petersen, T.N., Brunak, S., von Heijne, G. & Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10), pp.785-6.
- Peuvel, I., Delbac, F., Metenier, G., Peyret, P. & Vivares, C.P., 2000. Polymorphism of the gene encoding a major polar tube protein PTP1 in two microsporidia of the genus *Encephalitozoon*. *Parasitology*, 121(6), pp.581-7.
- Peyretailade, E., Broussolle, V., Peyret, P., Méténier, G., Gouy, M. & Vivares, C.P., 1998. Microsporidia, amitochondrial protists, possess a 70-kDa heat shock protein gene of mitochondrial evolutionary origin. *Molecular biology and evolution*, 15(6), pp.683-9.
- Peyretailade, E., Gonçalves, O., Terrat, S., Dugat-Bony, E., Wincker, P., Cornman, R.S., Evans, J.D., Delbac, F. & Peyret, P., 2009. Identification of transcriptional signals in *Encephalitozoon cuniculi* widespread among Microsporidia phylum: support for accurate structural genome annotation. *BioMed central genomics*, 10, p.607.

Peyretailade, E., Parisot, N., Polonais, V., Terrat, S., Denonfoux, J., Dugat-Bony, E., Wawrzyniak, I., Biderre-Petit, C., Mahul, A., Rimour, S. & Gonçalves, O., 2012. Annotation of microsporidian genomes using transcriptional signals. *Nature communications*, 3, p.1137.

Polonais, V., Mazet, M., Wawrzyniak, I., Texier, C., Blot, N., El Alaoui, H. & Delbac, F., 2010. The human microsporidian *Encephalitozoon hellem* synthesizes two spore wall polymorphic proteins useful for epidemiological studies. *Infection and immunity*, 78(5), pp.2221-30.

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).

Pombert, J.F., Selman, M., Burki, F., Bardell, F.T., Farinelli, L., Solter, L.F., Whitman, D.W., Weiss, L.M., Corradi, N. & Keeling, P.J., 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(31), pp.12638-43.

Pombert, J.F., Xu, J., Smith, D.R., Heiman, D., Young, S., Cuomo, C.A., Weiss, L.M. & Keeling, P.J., 2013. Complete genome sequences from three genetically distinct strains reveal high intraspecies genetic diversity in the microsporidian *Encephalitozoon cuniculi*. *Eukaryotic cell*, 12(4), pp.503-11.

Pongponratn, E., Maneerat, Y., Chaisri, U., Wilairatana, P., Punpoowong, B., Viriyavejakul, P. & Riganti, M., 1998. Electron-microscopic examination of *Rickettsia tsutsugamushi*-infected human liver. *Tropical medicine & international health*, 3(3), pp.242–8.

Pongsri, C. & Sukumasavin, N., 2005. National Aquaculture Sector Overview. Thailand. National Aquaculture Sector Overview Fact Sheets. *FAO Fisheries and Aquaculture Department*. Available at: http://www.fao.org/fishery/countrysector/naso_thailand/en [Accessed January 31, 2016].

Rabodonirina, M., Bertocchi, M., Desportes-Livage, I., Cotte, L., Levrey, H., Piens, M.A., Monneret, G., Celard, M., Mornex, J.F. & Mojon, M., 1996.

Enterocytozoon bieneusi as a cause of chronic diarrhea in a heart-lung transplant recipient who was seronegative for human immunodeficiency virus. *Clinical infectious diseases*, 23(1), pp.114-7.

Ramel, A.H., Barnard, E.A., Rustum, Y.M. & Jones, J.G., 1971. Yeast Hexokinase. IV. Multiple forms of hexokinase in the yeast cell. *Biochemistry*, 10(19), pp.3499-508.

Ramsay, K., Kaiser, M.J., Moore, P.G. & Hughes, R.N., 1997. Consumption of fisheries discards by benthic scavengers: utilization of energy subsidies in different marine habitats. *Journal of animal ecology*, 66(6), pp.884-896.

Rausch, M. & Grunewald, J., 1980. Light and stereoscan electron microscopic observations on some Microsporidian parasites (Cnidosporidia: Microsporidia) of blackfly larvae (Diptera: Simuliidae). *Parasitology research*, 63(1), pp.1-11.

Reetz, J., Nöckler, K., Reckinger, S., Vargas, M.M., Weiske, W. & Broglia, A., 2009. Identification of *Encephalitozoon cuniculi* genotype III and two novel genotypes of *Enterocytozoon bieneusi* in swine. *Parasitology international*, 58(3), pp.285-92.

Reetz, J., Rinder, H., Thomschke, A., Manke, H., Schwebs, M. & Bruderek, A., 2002. First detection of the microsporidium *Enterocytozoon bieneusi* in non-mammalian hosts (chickens). *International journal for parasitology*, 32(7), pp.785-7.

Reeves, R.E., 1968. A new enzyme with the glycolytic function of pyruvate kinase. *The Journal of biological chemistry*, 243(11), pp.3202-4.

Reeves, R.E., Montalvo, F. & Sillero, A., 1967. Glucokinase from *Entamoeba histolytica* and related organisms. *Biochemistry*, 6(6), pp.1752-60.

Reite, O.B., 2005. The rodlet cells of teleostean fish: their potential role in host defence in relation to the role of mast cells/eosinophilic granule cells. *Fish and shellfish immunology*, 19(3), pp.253-67.

Renesto, P., Ogata, H., Audic, S., Claverie, J.M. & Raoult, D., 2005. Some lessons from *Rickettsia* genomics. *Federation of European microbiological*

societies microbiology reviews, 29(1), pp.99-117.

Rhoads, A. & Au, K.F., 2015. PacBio Sequencing and Its applications. *Genomics, proteomics and bioinformatics*, 13(5), pp.278-289.

Rice, P., Longden, I. & Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. *Trends in genetics*, 16(6), pp.276-7.

Richards, T.A., Hirt, R.P., Williams, B.A. & Embley, T.M., 2003. Horizontal gene transfer and the evolution of parasitic protozoa. *Protist*, 154(1), pp.17-32.

Rocha, E.P.C. & Danchin, A., 2002. Base composition bias might result from competition for metabolic resources. *Trends in genetics*, 18(6), pp.291-4.

Rodríguez, A., Herrero, P. & Moreno, F., 2001. The hexokinase 2 protein regulates the expression of the GLK1, HXK1 and HXK2 genes of *Saccharomyces cerevisiae*. *The Biochemical journal*, 355(3), pp.625-31.

Roger, A.J., Clark, C.G. & Doolittle, W.F., 1996. A possible mitochondrial gene in the early-branching amitochondriate protist *Trichomonas vaginalis*. *Proceedings of the national academy of sciences of the United States of America*, 93(25), pp.14618-22.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A. & Huelsenbeck, J.P., 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3), pp. 539-42

Roudel, M., Aufauvre, J., Corbara, B., Delbac, F. & Blot, N., 2013. New insights on the genetic diversity of the honeybee parasite *Nosema ceranae* based on multilocus sequence analysis. *Parasitology*, 140(11), pp.1346-56.

Russell, D.W., 1992. Cholesterol biosynthesis and metabolism. *Cardiovascular Drugs and Therapy*, 6(2), pp.103-110.

Saier, M.H., Tran, C. V & Barabote, R.D., 2006. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic acids research*, 34, pp.D181-6.

Sak, B., Kvác, M., Petrzeková, K., Kvetonová, D., Pomajbíková, K., Mulama, M., Kiyang, J. & Modrý, D., 2011. Diversity of microsporidia (Fungi: Microsporidia) among captive great apes in European zoos and African sanctuaries: evidence for zoonotic transmission? *Folia parasitologica*, 58(2), pp.81-6.

Salim, H.M., Koire, A.M., Stover, N.A. & Cavalcanti, A.R., 2011. Detection of fused genes in eukaryotic genomes using gene deFuser: analysis of the *Tetrahymena thermophila* genome. *BioMed central bioinformatics*, 12(1), p.279.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M. & Marçais, G., 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome research*, 22(3), pp.557-67.

Sanger, F. & Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3), pp.441-448.

Sant'Anna, B.S., Santos, D.M., Marchi, M.R.R.D., Zara, F.J. & Turra, A., 2014. Surface-sediment and hermit-crab contamination by butyltins in southeastern Atlantic estuaries after ban of TBT-based antifouling paints. *Environmental science and pollution research international*, 21(10), pp.6516-24.

Santín, M. & Fayer, R., 2011. Microsporidiosis: *Enterocytozoon bieneusi* in domesticated and wild animals. *Research in veterinary science*, 90(3), pp.363-71.

Sax, P.E., Rich, J.D., Pieciak, W.S. & Trnka, Y.M., 1995. Intestinal microsporidiosis occurring in a liver transplant recipient. *Transplantation*, 60, pp.617-8.

Scanlon, M., Leitch, G.J., Visvesvara, G.S. & Shaw, A.P., 2004. Relationship between the host cell mitochondria and the parasitophorous vacuole in cells infected with *Encephalitozoon* Microsporidia. *The Journal of eukaryotic microbiology*, 51(1), pp.81-87.

Schilling, C.H., Covert, M.W., Famili, I., Church, G.M., Edwards, J.S. &

Palsson, B.O., 2002. Genome-scale metabolic model of *Helicobacter pylori* 26695. *Journal of bacteriology*, 184(16), pp.4582-93.

Schmieder, R. & Edwards, R., 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *Public library of science one*, 6(3).

Schnarrenberger, C., 1990. Characterization and compartmentation, in green leaves, of hexokinases with different specificities for glucose, fructose, and mannose and for nucleoside triphosphates. *Planta*, 181(2), pp.249-255.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Bolchacova, E., Voigt, K., Crous, P.W. & Miller, A.N., 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the national academy of sciences of the United States of America*, 109(16), pp.6241-6.

Schulz, F. & Horn, M., 2015. Intranuclear bacteria: inside the cellular control center of eukaryotes. *Trends in Cell Biology*, 25(6), pp.339–46.

Schulz, F., Lagkouvardos, I., Wascher, F., Aistleitner, K., Kostanjšek, R. & Horn, M., 2014. Life in an unusual intracellular niche: a bacterial symbiont infecting the nucleus of amoebae. *The ISME Journal*, 8(8), pp.1634–44.

Seafish, 2015. UK landings. Available at: http://www.seafish.org/media/publications/Seafood_Industry_Factsheet_2015.pdf [Accessed January 30, 2016].

Senderskiy, I.V., Timofeev, S.A., Seliverstova, E.V., Pavlova, O.A. & Dolgikh, V.V., 2014. Secretion of *Antonospora (Paranosema) locustae* proteins into infected cells suggests an active role of microsporidia in the control of host programs and metabolic processes. *Public library of science one*, 9(4).

Sewankambo, N.K., Gray, R.H., Ahmad, S., Serwadda, D., Wabwire-Mangen, F., Nalugoda, F., Kiwanuka, N., Lutalo, T., Kigozi, G., Li, C. & Meehan, M.P., 2000. Mortality associated with HIV infection in rural Rakai District, Uganda. *AIDS*, 14(15), pp.2391-400.

Shadduck, J.A. & Pakes, S.P., 1971. Encephalitozoonosis (nosematosis) and

- toxoplasmosis. *The American journal of pathology*, 64(3), pp.657-72.
- Shadduck, J.A., Meccoli, R.A., Davis, R. & Font, R.L., 1990. Isolation of a microsporidian from a human patient. *The Journal of infectious diseases*, 162(3), pp.773-6.
- Shendure, J. & Ji, H., 2008. Next-generation DNA sequencing. *Nature biotechnology*, 26(10), pp.1135-45.
- Sibly, R.M. & McCleery, R.H., 1983. The Distribution between feeding sites of herring gulls breeding at Walney Island, U.K. *The journal of animal ecology*, 52(1), p.51.
- Sikora, A.E., Zielke, R.A., Lawrence, D.A., Andrews, P.C. & Sandkvist, M., 2011. Proteomic analysis of the *Vibrio cholerae* type II secretome reveals new proteins, including three related serine proteases. *Journal of Biological Chemistry*, 286(19), pp.16555-66.
- Silveira, H. & Canning, E.U., 1995. *Vittaforma corneae* n. comb. for the human microsporidium *Nosema corneum* Shadduck, Meccoli, Davis & Font, 1990, based on its ultrastructure in the liver of experimentally infected athymic mice. *The Journal of eukaryotic microbiology*, 42(2), pp.158-65.
- Silverman, J.M., Chan, S.K., Robinson, D.P., Dwyer, D.M., Nandan, D., Foster, L.J. & Reiner, N.E., 2008. Proteomic analysis of the secretome of *Leishmania donovani*. *Genome Biology*, 9(2), p.R35.
- Silverman, J.M., Chan, S.K., Robinson, D.P., Dwyer, D.M., Nandan, D., Foster, L.J. & Reiner, N.E., 2004. Genome compaction and stability in microsporidian intracellular parasites. *Current biology*, 14(10), pp.891-6.
- Simpson, G.G., 1961. *Principles of Animal Taxonomy*, New York: Columbia University Press.
- Sinai, A.P., Webster, P. & Joiner, K.A., 1997. Association of host cell endoplasmic reticulum and mitochondria with the *Toxoplasma gondii* parasitophorous vacuole membrane: a high affinity interaction. *Journal of cell science*, (17), pp.2117-28.
- Skube, S.B., Chaverri, J.M. & Goodson, H. V., 2010. Effect of GFP tags on

the localization of EB1 and EB1 fragments *in vivo*. *Cytoskeleton*, 67(1), pp.1-12.

Slamovits, C.H. & Keeling, P.J., 2004. Class II photolyase in a microsporidian intracellular parasite. *Journal of molecular biology*, 341(3), pp.713-21.

Slamovits, C.H., Williams, B.A.P. & Keeling, P.J., 2004. Transfer of *Nosema locustae* (Microsporidia) to *Antonospora locustae* n. comb. Based on molecular and ultrastructural data. *The Journal of eukaryotic microbiology*, 51(2), pp.207-213.

Slodkowicz-Kowalska, A., Graczyk, T.K., Tamang, L., Jedrzejewski, S., Nowosad, A., Zduniak, P., Solarczyk, P., Girouard, A.S. & Majewska, A.C., 2006. Microsporidian species known to infect humans are present in aquatic birds: implications for transmission via water? *Applied and environmental microbiology*, 72(7), pp.4540-4.

Snowden, K., Logan, K. & Didier, E.S., 1999. *Encephalitozoon cuniculi* strain III is a cause of encephalitozoonosis in both humans and dogs. *The Journal of infectious diseases*, 180(6), pp.2086-8.

Söderhäll, K. & Thörnqvist, P.O., 1997. Crustacean immunity-a short review. *Developments in biological standardization*, 90, pp.45-51.

Sokolova, Y.Y., Dolgikh, V.V., Morzhina, E.V., Nassonova, E.S., Issi, I.V., Terry, R.S., Ironside, J.E., Smith, J.E. & Vossbrinck, C.R., 2003. Establishment of the new genus *Paranosema* based on the ultrastructure and molecular phylogeny of the type species *Paranosema grylli* gen. nov., comb. nov. (Sokolova, Seleznirov, Dolgikh, Issi 1994), from the cricket *Gryllus bimaculatus* Deg. *Journal of invertebrate pathology*, 84(3), pp.159-72.

Sokolova, Y.Y., Fuxa, J.R. & Borkhsenius, O.N., 2005. The nature of *Thelohania solenopsae* (Microsporidia) cysts in abdomens of red imported fire ants, *Solenopsis invicta*. *Journal of invertebrate pathology*, 90(1), pp.24-31.

Sokolova, Y.Y., Lange, C.E., Mariottini, Y. & Fuxa, J.R., 2009. Morphology and taxonomy of the microsporidium *Liebermannia covasacrae* n. sp. from the grasshopper *Covasacris pallidinota* (Orthoptera, Acrididae). *Journal of invertebrate pathology*, 101(1), pp.34-42.

Solis-Lucero, G., Manoutcharian, K., Hernández-López, J. & Ascencio, F., 2016. Injected phage-displayed-VP28 vaccine reduces shrimp *Litopenaeus vannamei* mortality by white spot syndrome virus infection. *Fish & Shellfish Immunology*, 55, pp.401-406.

Sprague, V. & Becnel, J.J., 1998. Note on the Name-Author-Date combination for the taxon *Microsporidies balbiani*, 1882, when ranked as a phylum. *Journal of invertebrate pathology*, 71(1), pp.91-94.

Sprague, V. & Vernick, S.H., 1969. Light and electron microscope observations on *Nosema nelsoni* Sprague, 1950 (Microsporida, Nosematidae) with particular reference to its golgi complex. *The Journal of protozoology*, 16(2), pp.264-271.

Sprague, V., 1977. *Systematics of the Microsporidia* 2nd ed. L. A. J. Bulla & T. Cheng, eds., New York: Plenum Press.

Sprague, V., Becnel, J.J. & Hazard, E.I., 1992. Taxonomy of phylum microspora. *Critical reviews in microbiology*, 18(5-6), pp.285-395.

Sprague, V., Ormieres, R. & Manier, J.F., 1972. Creation of a new genus and a new family in the Microsporida. *Journal of invertebrate pathology*, 20(2), pp.228-231.

Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), pp.1312-3.

Stark, D., Van Hal, S., Barratt, J., Ellis, J., Marriott, D. & Harkness, J., 2009. Limited genetic diversity among genotypes of *Enterocytozoon bieneusi* strains isolated from HIV-infected patients from Sydney, Australia. *Journal of medical microbiology*, 58(3), pp.355-7.

Stechmann, A., Baumgartner, M., Silberman, J.D. & Roger, A.J., 2006. The glycolytic pathway of *Trimastix pyriformis* is an evolutionary mosaic. *BioMed central evolutionary biology*, 6(1), p.101.

Stempell, W., 1909. Uber *Nosema bombycis* Nageli. *Arch. Protistenkdrotis*, 16, pp.281-358.

Stentiford, G.D. & Bateman, K.S., 2007. *Enterospora* sp ., an intranuclear

microsporidian infection of hermit crab *Eupagurus bernhardus*. *Diseases of aquatic organisms*, 75, pp.73-78.

Stentiford, G.D., Bateman, K.S., Dubuffet, A., Chambers, E. & Stone, D.M., 2011. *Hepatospora eriocheir* (Wang and Chen, 2007) gen. et comb. nov. infecting invasive Chinese mitten crabs (*Eriocheir sinensis*) in Europe. *Journal of invertebrate pathology*, 108(3), pp.156-66.

Stentiford, G.D., Bateman, K.S., Feist, S.W., Chambers, E. & Stone, D.M., 2013. Plastic parasites: extreme dimorphism creates a taxonomic conundrum in the phylum Microsporidia. *International journal for parasitology*, 43(5), pp.339-52.

Stentiford, G.D., Bateman, K.S., Longshaw, M. & Feist, S.W. 2007. *Enterospora canceri* n. gen., n. sp., intranuclear within the hepatopancreatocytes of the European edible crab *Cancer pagurus*. *Diseases of aquatic organisms*, 75(1), pp.61-72.

Stentiford, G.D., Bateman, K.S., Small, H.J., Pond, M. & Ungfors, A., 2012. *Hematodinium* sp. and its bacteria-like endosymbiont in European brown shrimp (*Crangon crangon*). *Aquatic biosystems*, 8(1), p.24.

Stentiford, G.D., Becnel, J.J., Weiss, L.M., Keeling, P.J., Didier, E.S., Williams, B.A.P., Bjornson, S., Kent, M.L., Freeman, M.A., Brown, M.J.F. & Troemel, E.R., 2016. Microsporidia - Emergent pathogens in the global food chain. *Trends in parasitology*, 32(4), pp.336-48.

Stentiford, G.D., Feist, S.W., Stone, D.M., Bateman, K.S. & Dunn, A.M., 2013. Microsporidia: diverse, dynamic, and emergent pathogens in aquatic systems. *Trends in parasitology*, 29(11), pp.567-78.

Stentiford, G.D., Feist, S.W., Stone, D.M., Peeler, E.J. & Bass, D., 2014. Policy, phylogeny, and the parasite. *Trends in parasitology*, 30(6), pp.274-81.

Stiller, J.W. & Hall, B.D., 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. *Molecular biology and evolution*, 16(9), pp.1270-9.

Stoltzfus, A., 1999. On the possibility of constructive neutral evolution. *Journal of Molecular Evolution*, 49(2), pp.169-181.

Suga, H., Koyanagi, M., Hoshiyama, D., Ono, K., Iwabe, N., Kuma, K.I. & Miyata, T., 1999. Extensive gene duplication in the early evolution of animals before the parazoan-eumetazoan split demonstrated by G proteins and protein tyrosine kinases from sponge and hydra. *Journal of molecular evolution*, 48(6), pp.646-53.

Sulaiman, I.M., Fayer, R., Lal, A.A., Trout, J.M., Schaefer, F.W. & Xiao, L., 2003. Molecular characterization of microsporidia indicates that wild mammals Harbor host-adapted *Enterocytozoon* spp. as well as human-pathogenic *Enterocytozoon bieneusi*. *Applied and environmental microbiology*, 69(8), pp.4495-501.

Sulaiman, I.M., Fayer, R., Yang, C., Santin, M., Matos, O. & Xiao, L., 2004. Molecular characterization of *Enterocytozoon bieneusi* in cattle indicates that only some isolates have zoonotic potential. *Parasitology research*, 92(4), pp.328-34.

Sweet, M.J. & Bateman, K.S., 2015. Diseases in marine invertebrates associated with mariculture and commercial fisheries. *Journal of sea research*, 104, pp.16-32.

Takeuchi, H., Hirano, T., Whitmore, S.E., Morisaki, I., Amano, A. & Lamont, R.J., 2013. The serine phosphatase SerB of *Porphyromonas gingivalis* suppresses IL-8 production by dephosphorylation of NF- κ B RelA/p65. *Public library of science pathogens*, 9(4).

Takizawa, H., Vivier, E. & Petitprez, A., 1973. Développement intranucléaire de la Microsporidie *Nosema bombycis* dans les cellules de vers à soie après infestation expérimentale. *C. R. Acad. Sci.*, 277, pp.1769-72.

Takvorian, P.M. & Cali, A., 1994. Enzyme histochemical identification of the Golgi apparatus in the microsporidian, *Glugea stephani*. *The Journal of eukaryotic microbiology*, 41(5), p.63S-64S.

Takvorian, P.M., Weiss, L.M. & Cali, A., 2005. The early events of *Brachiola algerae* (Microsporidia) infection: spore germination, sporoplasm structure, and development within host cells. *Folia parasitologica*, 52(1-2), pp.118-29.

Tangprasittipap, A., Srisala, J., Chouwdee, S., Somboon, M., Chuchird, N.,

Limsuwan, C., Srisuvan, T., Flegel, T.W. & Sritunyalucksana, K., 2013. The microsporidian *Enterocytozoon hepatopenaei* is not the cause of white feces syndrome in whiteleg shrimp *Penaeus (Litopenaeus) vannamei*. *BioMed central veterinary research*, 9, p.139.

Tarailo-Graovac, M. & Chen, N., 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, pp.4-10.

Tatusova, T., Ciufu, S., Fedorov, B., O'Neill, K. & Tolstoy, I., 2014. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic acids research*, 42, pp.D553-9.

Taupin, V., Garenaux, E., Mazet, M., Maes, E., Denise, H., Prensier, G., Vivarès, C.P., Guérardel, Y. & Méténier, G., 2007. Major O-glycans in the spores of two microsporidian parasites are represented by unbranched manno-oligosaccharides containing alpha-1,2 linkages. *Glycobiology*, 17(1), pp.56-67.

Tian, S., Zhu, L. & Liu, M., 2010. Bioaccumulation and distribution of polybrominated diphenyl ethers in marine species from Bohai Bay, China. *Environmental toxicology and chemistry*, 29(10), pp.2278-85.

Tilney, L.G., Harb, O.S., Connelly, P.S., Robinson, C.G. & Roy, C.R., 2001. How the parasitic bacterium *Legionella pneumophila* modifies its phagosome and transforms it into rough ER: implications for conversion of plasma membrane to the ER membrane. *Journal of cell science*, 114(24), pp.4637-50.

Tokarev, Y., Sitnikova, N. & Pistone, D., 2010. Microsporidia PCR detection artifacts due to non-specific binding of the universal microsporidia primers to the rDNA of arthropod hosts. *Journal of academy of science of Moldova life sciences*, 310(1).

Tokarev, Y.S., Voronin, V.N., Seliverstova, E.V., Dolgikh, V.V., Pavlova, O.A., Ignatieva, A.N. & Issi, I.V., 2010. Ultrastructure and molecular phylogeny of *Anisofilariata chironomi* g.n. sp.n. (Microsporidia: Terresporidia) from *Chironomus plumosus* L. (Diptera: Chironomidae). *Parasitology research*,

107(1), pp.39-46.

Tort, L., Padros, F., Rotland, J. & Crespo, S., 1998. Winter syndrome in the gilthead sea bream *Sparus aurata*. Immunological and histopathological features. *Fish & shellfish immunology*, 8(1), pp.37-47.

Tort, L., Rotllant, J. & Rovira, L., 1998. Immunological suppression in gilthead sea bream *Sparus aurata* of the North-West Mediterranean at low temperatures. *Comparative Biochemistry and Physiology. Molecular & Integrative Physiology*, 120(1), pp.175-179.

Tourtip, S., Wongtripop, S., Stentiford, G.D., Bateman, K.S, Sriurairatana, S., Chavadej, J., Sritunyalucksana, K. & Withyachumnarnkul, B., 2009. *Enterocytozoon hepatopenaei* sp. nov. (Microsporida: Enterocytozoonidae), a parasite of the black tiger shrimp *Penaeus monodon* (Decapoda: Penaeidae): Fine structure and phylogenetic relationships. *Journal of invertebrate pathology*, 102(1), pp.21-9.

Tribble, G.D., Mao, S., James, C.E. & Lamont, R.J., 2006. A *Porphyromonas gingivalis* haloacid dehalogenase family phosphatase interacts with human phosphoproteins and is important for invasion. *Proceedings of the National Academy of Sciences*, 103(29), pp.11027-11032.

Tritt, A., Eisen, J.A., Facciotti, M.T. & Darling, A.E., 2012. An integrated pipeline for *de novo* assembly of microbial genomes. *Public library of science one*, 7(9).

Troemel, E.R., Félix, M.A., Whiteman, N.K., Barrière, A. & Ausubel, F.M., 2008. Microsporidia are natural intracellular parasites of the nematode *Caenorhabditis elegans*. *Public library of science biology*, 6(12).

Tsang, L.M., Schubart, C.D., Ahyong, S.T., Lai, J.C., Au, E.Y., Chan, T.Y., Ng, P.K. & Chu, K.H., 2014. Evolutionary history of true crabs (Crustacea: Decapoda: Brachyura) and the origin of freshwater crabs. *Molecular biology and evolution*, 31(5), pp.1173-87.

Tsaousis, A.D., Kunji, E.R., Goldberg, A.V., Lucocq, J.M., Hirt, R.P. & Embley, T.M., 2008. A novel route for ATP acquisition by the remnant mitochondria of *Encephalitozoon cuniculi*. *Nature*, 453(7194), pp.553-6.

Tumwine, J.K., Kekitiinwa, A., Bakeera-Kitaka, S., Ndeezi, G., Downing, R., Feng, X., Akiyoshi, D.E. & Tzipori, S., 2005. Cryptosporidiosis and microsporidiosis in ugandan children with persistent diarrhea with and without concurrent infection with the human immunodeficiency virus. *The American journal of tropical medicine and hygiene*, 73(5), pp.921-5.

Tuzet, O., Maurand, J., Fize, A., Michel, R. & Fenwick, B., 1971. Proposition d'un nouveau cadre systématique pour les genres de Microsporidies. *Comptes rendus de l'Académie des Sciences.*, 272, pp.1268-1271.

Undeen, A.H. & Epsky, N.D., 1990. In vitro and in vivo germination of *Nosema locustae* (Microsporida: Nosematidae) spores. *Journal of invertebrate pathology*, 56, pp.371-379.

Undeen, A.H. & Vander Meer, R.K., 1994. Conversion of intrasporal trehalose into reducing sugars during germination of *Nosema algerae* (Protista: Microspora) spores: A quantitative study. *The journal of eukaryotic microbiology*, 41(2), pp.129-132.

Undeen, A.H. & Vander Meer, R.K., 1999. Microsporidian intrasporal sugars and their role in germination. *Journal of invertebrate pathology*, 73(3), pp.294-302.

Undeen, A.H., 1983. The germination of *Vavraia culicis* spores. *The Journal of Protozoology*, 30(2), pp.274-277.

Utturkar, S.M., Klingeman, D.M., Land, M.L., Schadt, C.W., Doktycz, M.J., Pelletier, D.A. & Brown, S.D., 2014. Evaluation and validation of *de novo* and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics*, 30(19), pp.2709-16.

Valdez, A., Yepiz-Plascencia, G., Ricca, E. & Olmos, J., 2014. First *Litopenaeus vannamei* WSSV 100% oral vaccination protection using CotC::Vp26 fusion protein displayed on *Bacillus subtilis* spores surface. *Journal of applied microbiology*, 117(2), pp.347-57.

Van der Zee, R., Gómez-Moracho, T., Pisa, L., Sagastume, S., García-Palencia, P., Maside, X., Bartolomé, C., Martín-Hernández, R. & Higes, M.,

2014. Virulence and polar tube protein genetic diversity of *Nosema ceranae* (Microsporidia) field isolates from Northern and Southern Europe in honeybees (*Apis mellifera iberiensis*). *Environmental microbiology reports*, 6(4), pp.401-13.

van Hoof, A., 2005. Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. *Genetics*, 171(4), pp.1455-61.

Vannini, M. & Cannicci, S., 1995. Homing behaviour and possible cognitive maps in crustacean decapods. *Journal of experimental marine biology and ecology*, 193(1-2), pp.67-91.

Vávra, J. & Larson, J.I.R., 1999. Structure of the microsporidia. In L. M. Weiss & M. Wittner, eds. *The Microsporidia and Microsporidiosis*. Washington D. C.: ASM Press, p. 55.

Vávra, J. & Larson, J.I.R., 2014. Structure of Microsporidia. In L. M. Weiss & J. J. Becnel, eds. *Microsporidia: Pathogens of opportunity*. John Wiley and Sons, inc., pp. 1-70.

Vávra, J. & Lukeš, J., 2013. Microsporidia and 'the art of living together'. *Advanced parasitology*, 82, pp.253-319.

Vávra, J. & Undeen, A.H., 1970. *Nosema algerae* n. sp. (Cnidospora, Microsporida) a pathogen in a laboratory colony of *Anopheles stephensi* Liston (Diptera, Culicidae). *The Journal of protozoology*, 17(2), pp.240-249.

Vávra, J., 1965. Study by electron microscope of the morphology and development of some Microsporidia. *Sciences naturelles*, 261(17), pp.3467-70.

Vávra, J., 1976. Biology of the Microsporidia. In L. A. Bulla & T. C. Cheng, eds. *Comparative pathobiology*. New York: Springer, pp. 1-86.

Vávra, J., 1977. Structure of the microsporidia. In L. A. Bulla & T. C. Cheng eds. *Comparative pathobiology*. New York: Plenum Press, pp. 1-85.

Vávra, J., Vinckier, D., Torpier, G., Porchet, E. & Vivier, E., 1986. A freeze-fracture study of microsporidia (Protozoa: Microspora). I. The sporophorous

vesicle, the spore wall, the spore plasma membrane. *Parasitologica*, 22, pp.143-154.

Vivarès, C.P., Gouy, M., Thomarat, F. & Méténier, G., 2002. Functional and evolutionary analysis of a eukaryotic parasitic genome. *Current opinion in microbiology*, 5(5), pp.499-505.

Vivier, E., 1979. Données nouvelles sur les Microsporidies. Ultrastructure-cycles-systématique. *Bulletin de la Société zoologique de France*, 104, pp.381-396.

Voronin, V.N., 2001. On a macrotaxonomy of the phylum Microsporidia. *Parasitologiya*, 35, pp.35-34.

Vossbrinck, C.R. & Debrunner-Vossbrinck, B. A., 2005. Molecular phylogeny of the Microsporidia: ecological, ultrastructural and taxonomic considerations. *Folia parasitologica*, 52(1), pp.131-142.

Vossbrinck, C.R. & Woese, C.R., 1986. Eukaryotic ribosomes that lack a 5.8S RNA. *Nature*, 320(6059), pp.287-288.

Vossbrinck, C.R., Andreadis, T.G. & Debrunner-Vossbrinck, B.A., 1998. Verification of intermediate hosts in the life cycles of Microsporidia by small subunit rDNA sequencing. *The Journal of Eukaryotic Microbiology*, 45(3), pp.290-292.

Vossbrinck, C.R., Baker, M.D., Didier, E.S., Debrunner-Vossbrinck, B.A. & Shadduck, J.A. 1993. Ribosomal DNA Sequences of *Encephalitozoon hellem* and *Encephalitozoon cuniculi*: Species identification and phylogenetic construction. *The Journal of eukaryotic microbiology*, 40(3), pp.354-362.

Vossbrinck, C.R., Maddox, J.V., Friedman, S., Debrunner-Vossbrinck, B.A. & Woese, C.R., 1987. Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. *Nature*, 326(6111), pp.411-4.

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. *Public library of science one*, 9(11).

Wang, W. & Chen, J., 2007. Ultrastructural study on a novel microsporidian, *Endoreticulatus eriocheir* sp. nov. (Microsporidia, *Encephalitozoonidae*), parasite of Chinese mitten crab, *Eriocheir sinensis* (Crustacea, Decapoda). *Journal of invertebrate pathology*, 94(2), pp.77-83.

Wang, W. & Gu, Z., 2002. *Rickettsia*-like organism associated with tremor disease and mortality of the Chinese mitten crab *Eriocheir sinensis*. *Diseases of aquatic organisms*, 48(2), pp.149-53.

Watson, A.K., Williams, T.A., Williams, B.A., Moore, K.A., Hirt, R.P. & Embley, T.M., 2015. Transcriptomic profiling of host-parasite interactions in the microsporidian *Trachipleistophora hominis*. *BioMed central genomics*, 16(1), p.1.

Watts, M.R., Chan, R.C., Cheong, E.Y., Brammah, S., Clezy, K.R., Tong, C., Marriott, D., Webb, C.E., Chacko, B., Tobias, V. & Outhred, A.C., 2014. *Anncaliia algerae* microsporidial myositis. *Emerging infectious diseases*, 20(2), p.185.

Weber, R., Müller, A., Spycher, M.A., Opravil, M., Ammann, R. & Briner, J., 1992. Intestinal *Enterocytozoon bieneusi* microsporidiosis in an HIV-infected patient: diagnosis by ileocolonoscopy biopsies and long-term follow up. *The clinical investigator*, 70(11), pp.1019-23.

Weidner, E. & Byrd, W., 1982. The Microsporidian spore invasion tube-Role of calcium in the activation of invasion tube discharge. *The journal of cell biology*, 93, pp.970-975.

Weidner, E., Byrd, W., Scarborough, A., Pleshinger, J. & Sibley, D., 1984. Microsporidian spore discharge and the transfer of polaroplast organelle membrane into plasma membrane. *The Journal of Protozoology*, 31(2), pp.195-198.

Weimin, M., 2006. Cultured Aquatic Species Information Programme. *Eriocheir sinensis*. *FAO Fisheries and Aquaculture Department*. Available at: http://www.fao.org/fishery/culturedspecies/Eriocheir_sinensis/en#tcNA00EA [Accessed January 31, 2016].

Weiser, J., 1977. Contribution to the classification of Microsporidia. *Vestnik*

Cesl Spol Zoo, 41(4), pp.308-321.

Weissenberg, R., 1976. Microsporidian interactions with host cells. In J. L. A. Bulla & T. C. Cheng, eds. *Comparative Pathobiology*. New York: Plenum Press, pp. 203-237.

Wernegreen, J.J., Lazarus, A.B. & Degnan, P.H., 2002. Small genome of *Candidatus blochmannia*, the bacterial endosymbiont of *Camponotus*, implies irreversible specialization to an intracellular lifestyle. *Microbiology*, 148(8), pp.2551-2556.

Wernersson, R., 2005. FeatureExtract-extraction of sequence annotation made easy. *Nucleic acids research*, 33, pp.W567-9.

Werren, J.H., Baldo, L. & Clark, M.E., 2008. *Wolbachia*: master manipulators of invertebrate biology. *Nature reviews microbiology*, 6(10), pp.741-51.

Whitlock, V.H. & Johnson, S., 1990. Stimuli for the in vitro germination and inhibition of *Nosema locusta* (Microspora: Nosematidae) spores. *Journal of invertebrate pathology*, 56(1), pp.57-62.

Wichro, E., Hoelzl, D., Krause, R., Bertha, G., Reinthaler, F. & Wensch, C., 2005. Microsporidiosis in travel-associated chronic diarrhea in immune-competent patients. *The American journal of tropical medicine and hygiene*, 73(2), pp.285-7.

Widmer, G. & Akiyoshi, D.E., 2010. Host-specific segregation of ribosomal nucleotide sequence diversity in the microsporidian *Enterocytozoon bieneusi* infection, genetics and evolution. *Journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 10(1), pp.122-8.

Williams, B.A.P., Dolgikh, V. V. & Sokolova, Y.Y., 2014. Microsporidian Biochemistry and Physiology. In L. M. Weiss & J. J. Becnel, eds. *Microsporidia: Pathogens of opportunity*. John Wiley and Sons, inc., pp. 245-260.

Williams, B.A.P., Elliot, C., Burri, L., Kido, Y., Kita, K., Moore, A.L. & Keeling, P.J., 2010. A broad distribution of the alternative oxidase in microsporidian parasites. *Public library of science pathogens*, 6(2).

- Williams, B.A.P., Haferkamp, I. & Keeling, P.J., 2008. An ADP/ATP-specific mitochondrial carrier protein in the microsporidian *Antonospora locustae*. *Journal of molecular biology*, 375(5), pp.1249-57.
- Williams, B.A.P., Hirt, R.P., Lucocq, J.M. & Embley, T.M., 2002. A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*. *Nature*, 418(6900), pp.865-869.
- Williams, B.A.P., Lee, R.C., Becnel, J.J., Weiss, L.M., Fast, N.M. & Keeling, P.J., 2008. Genome sequence surveys of *Brachiola algerae* and *Edhazardia aedis* reveal microsporidia with low gene densities. *BioMed central genomics*, 9(1), p.200.
- Winde, J.H., Crauwels, M., Hohmann, S., Thevelein, J.M. & Winderickx, J., 1996. Differential requirement of the yeast sugar kinases for sugar sensing in establishing the catabolite-repressed state. *European journal of biochemistry*, 241(2), pp.633-43.
- Winters, A.D. & Faisal, M., 2014. Molecular and ultrastructural characterization of *Dictyocoela diporeiae* n. sp. (Microsporidia), a parasite of *Diporeia* spp. (Amphipoda, Gammaridea). *Parasite*, 21, p.26.
- Wiser, M.F., 2011. *Protozoa and Human Disease*, Garland Science.
- Wittner M. W. & Weiss, L.M., 1999. *The Microsporidia and Microsporidiosis*, Washington D. C.: ASM Press.
- Wolf, A., Agnihotri, S., Micallef, J., Mukherjee, J., Sabha, N., Cairns, R., Hawkins, C. & Guha, A., 2011. Hexokinase 2 is a key mediator of aerobic glycolysis and promotes tumor growth in human glioblastoma multiforme. *The Journal of experimental medicine*, 208(2), pp.313-326.
- Woolfit, M. & Bromham, L., 2003. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Molecular biology and evolution*, 20(9), pp.1545-55.
- Wright, N.A. & Irwin, M., 1982. The kinetics of villus cell populations in the mouse small intestine. I. Normal villi: the steady state requirement. *Cell and tissue kinetics*, 15(6), pp.595-609.

- Wu, G., Henze, K. & Müller, M., 2001. Evolutionary relationships of the glucokinase from the amitochondriate protist, *Trichomonas vaginalis*. *Gene*, 264(2), pp.265-271.
- Wu, M., Sun, L.V., Vamathevan, J., Riegler, M., Deboy, R., Brownlie, J.C., McGraw, E.A., Martin, W., Esser, C., Ahmadinejad, N. & Wiegand, 2004. Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile genetic elements. *Public library of science biology*, 2(3), p.E69.
- Xiang, H., Pan, G., Vossbrinck, C.R., Zhang, R., Xu, J., Li, T., Zhou, Z., Lu, C. & Xiang, Z., 2010. A tandem duplication of manganese superoxide dismutase in *Nosema bombycis* and its evolutionary origins. *Journal of molecular evolution*, 71(5-6), pp.401-14.
- Yagi, T. & Matsuno-Yagi, A., 2003. The Proton-Translocating NADH-Quinone Oxidoreductase in the Respiratory Chain: The Secret Unlocked. *Biochemistry*, 42(8), pp.2266-2274.
- Yang, J.Y., Chang, C.I., Liu, K.F., Hseu, J.R., Chen, L.H. & Tsai, J.M., 2012. Viral resistance and immune responses of the shrimp *Litopenaeus vannamei* vaccinated by two WSSV structural proteins. *Immunology letters*, 148(1), pp.41-48.
- Zerbino, D.R. & Birney, E., 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome research*, 18(5), pp.821-9.
- Zhang, Y. & Trissel, L.A., 2002. Stability of ampicillin sodium, nafcillin sodium, and oxacillin sodium in auto dose infusion system bags. *International journal of pharmaceutical compounding*, 6(3), pp.226-9.
- Zhao, W., Zhang, W., Yang, Z., Liu, A., Zhang, L., Yang, F., Wang, R. & Ling, H., 2015. Genotyping of *Enterocytozoon bieneusi* in farmed blue foxes (*Alopex lagopus*) and raccoon dogs (*Nyctereutes procyonoides*) in China. *Public library of science one*, 10(11).
- Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S. & Wang, W., 2008. On the origin of new genes in *Drosophila*. *Genome Research*, 18(9), pp.1446-1455.

Zhou, X. & Rokas, A., 2014. Prevention, diagnosis and treatment of high-throughput sequencing data pathologies. *Molecular ecology*, 23(7), pp.1679-700.

Zielinski, F.U., Pernthaler, A., Duperron, S., Raggi, L., Giere, O., Borowski, C. & Dubilier, N., 2009. Widespread occurrence of an intranuclear bacterial parasite in vent and seep bathymodiolin mussels. *Environmental Microbiology*, 11(5), pp.1150–67.