



Analyzing *Microsporidia* sp. MB
from Sequence to Biology: A
Comprehensive Exploration of the Genome,
Protein Structures, and Functions through
Extensive Bioinformatics Analysis

A thesis presented in fulfillment of the requirements for the degree
of Doctor of Philosophy in Bioinformatics

by

Ang'ang'o Lilian Mbaisi

<https://orcid.org/0000-0001-8573-5417>

Department of Biochemistry, Microbiology and Bioinformatics

Faculty of Science

Rhodes University, South Africa

September 21, 2024

Abstract

Microsporidia are spore-forming intracellular organisms classified as the earliest divergent group within the Fungi kingdom. Microsporidia have been found widely affecting different hosts, including both vertebrates and invertebrates. The pathogenicity of microsporidia depends on their species and the host species they infect. Due to their obligate intracellular nature, microsporidia have extensively evolved. This is illustrated by their highly variable genome sizes and gene content. Being minimalist eukaryotes, the microsporidia genome is often associated with extreme gene reduction and compaction. However, these interesting microorganisms retain particular genes that help them acquire specific host nutrients, thereby relying heavily on their host for survival and proliferation. The mode of sexual reproduction of microsporidia has not been well-studied. Harnessing microsporidia in the laboratory is often a challenge, however, the advances in computational tools have made it cheaper and quicker to accurately predict and annotate these organisms to understand their mechanism of infection. Understanding the protein structure and function of these unique organisms is the baseline for providing insights into their biology and survival in their respective hosts. Microsporidia genomes contain a large proportion of hypothetical proteins of which their functions are not described. *Vittaforma corneae* ATCC 50505 was used as a model to highlight the functions and structure of these otherwise unknown proteins. A systematic annotation pipeline employing exhaustive computational tools was devised to carefully annotate the hypothetical proteins of *V. corneae*, aiming to characterize their structure and function. The genome of the novel microsporidian, *Microsporidia* sp. MB, a *Plasmodium*-transmission-blocking symbiont isolated from *Anopheles* mosquitoes in Sub-Saharan Africa, was sequenced, assembled, and annotated. The genome was found to contain over 2000 putative genes spanning its

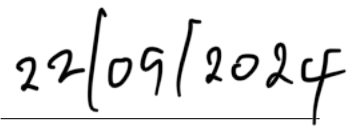
5.9 Mb size and contained minimal repeats. Comparative phylogenomic analysis of *Microsporidia* sp. MB grouped this symbiont within the Enterocytozoonida (clade IV) microsporidia, clustering with its closest relative – *V. corneae*. Using robust computational techniques, prediction and characterization of the putative proteins of *Microsporidia* sp. MB was conducted. The decay of several proteins in the glycolytic pathway is one unique characteristic associated with microsporidia. The proteins retained or lost often vary across the microsporidian taxon. This study highlights the retention of most of the proteins involved in the glycolytic pathway in *Microsporidia* sp. MB. The available genome dataset of *Microsporidia* sp. MB was further used to infer its mode of sexual reproduction. The symbiont appears to have several meiotic-related gene orthologs, suggesting that it is capable of sexual reproduction. These findings describe the basic biology of *Microsporidia* sp. MB and provide a basis for future Next-Generation Sequencing, RNA sequencing experiments ultimately informing the application of this microorganism as a biological malaria control tool.

Declaration

I declare that this thesis presented for the degree of Doctor of Philosophy in Science (Bioinformatics) for the Faculty of Science at Rhodes University has been solely composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.



ANG'ANG'O LILIAN MBAISI



DATE

Dedication

*To my parents David and Florence Ang'ang'o, and my siblings Elvis Ikoha and
Pauline Mwabishi, for everything.*

Acknowledgement

I am infinitely grateful to the Lord Almighty for the grace and strength throughout my Ph.D. journey. His comforting words of encouragement, patience, grace and love were my daily affirmations as I worked on this thesis.

I would like to sincerely express my gratitude to my supervisors Prof. Ozlem Tastan Bishop and Dr. Jeremy Keith Herren for their unwavering support and guidance throughout my research journey. Their expertise, encouragement, and valuable feedback have helped shape this research project.

My heartfelt gratitude goes to my parents, David and Florence Ang'ang'o for their motivation, financial support and guidance throughout this rigorous process. Their prayers kept me going. To my siblings, Pauline Mwabishi and Elvis Ikoha, thank you for making my role as a firstborn easy.

I am eternally grateful to my church family at The Cathedral of St. Michael and St. George, Makhandu for their unwavering support, spiritual growth, and selfless concerns throughout the global pandemic and after. I found a home away from home here. Wholehearted thanks to the Symbiovector Team at the International Center for Insect Physiology and Ecology (ICIPE) in Nairobi, Kenya for their support in collecting and preparing samples for sequencing. I would also like to acknowledge ICIPE for providing the invaluable resources to conduct these experiments.

My genuine gratitude goes to the Research Unit in Bioinformatics (RUBi) team for their support and feedback on my research. Their sincere curiosity in my work made me appreciate the impact of this work. I extend my heartfelt thanks to Rhodes University for their generous support in enabling me to pursue my research. Sincere thanks to the Centre for High Performance Computing (CHPC) for providing heavyweight computational resources to analyze the big data gener-

ated in this study.

Acknowledgement of Funding

My Ph.D. fellowship was supported by the **Organization for Women in Science for the Developing World (OWSD)**. This work was also supported by the **Bill & Melinda Gates Foundation (BMGF)**, **Open Philanthropy (OP)**, and the **Children's Investment Fund Foundation (CIFF)**.

Table of Contents

Abstract	ii
Declaration	iii
Dedication	iv
Acknowledgements	vi
List of Tables	xvi
List of Figures	xxviii
Acronyms/Abbreviations	xxxi
List of Abbreviations	xxxii
Research Outputs	xxxvii
List of Tools	xxxvii
Thesis Overview	xxxviii
1 Literature Overview	1
1.1 Background	1
1.1.1 The complex taxonomic classification of microsporidia . .	2
1.1.2 The microsporidia life cycle	5
1.1.3 Microsporidia genomes and genomic variations	7
1.1.4 Microsporidia in invertebrates	7

1.1.4.1	Development of new and innovative malaria control tools	8
1.1.4.2	The mosquito microbiome and impact on vectorial capacity	8
1.1.4.3	Microsporidia in vector control	9
1.1.4.3.1	Microsporidia of mosquitoes and their role in the control of vector-borne diseases	10
1.1.5	<i>Microsporidia</i> sp. MB as a potential vector control tool . .	12
1.2	Knowledge gap	14
1.3	Research aim and objectives	15
1.3.1	Research aim	15
1.3.2	Main objectives	15
2	Structural and Functional Annotation of Hypothetical Proteins from <i>Vittaforma corneae</i> ATCC 50505 Using in silico Approaches	17
2.1	Chapter Overview	17
2.2	Keywords	18
2.3	Introduction	18
2.3.1	<i>V. corneae</i> -host cell interactions	19
2.3.2	Hypothetical proteins (HPs)	19
2.3.3	Distribution of HPs in microsporidia proteomes	20
2.3.4	Computational approaches in protein functional and structural annotation	20
2.3.4.1	Protein databases	21
2.3.4.2	Protein homology search	21
2.3.4.3	Multiple sequence alignment (MSA) and phylogenetic analysis	22
2.3.4.4	Protein physicochemical analysis	23
2.3.4.5	Protein motif discovery	24
2.3.4.6	Conserved domains analysis	25
2.3.4.7	Homology modeling and three-dimensional (3D) structural analysis	26
2.4	Methodology	27

2.4.1	Phase I	29
2.4.1.1	Literature search of <i>V. corneae</i> proteins	29
2.4.1.2	Sequence retrieval	29
2.4.1.3	Homology search	30
2.4.2	Phase II	32
2.4.2.1	Protein family classification	32
2.4.2.2	Sub-cellular localization and signal peptide prediction	32
2.4.2.3	Analysis of protein physicochemical properties	33
2.4.2.4	KEGG and Gene Ontology analysis	33
2.4.3	Characterization of five case studies	33
2.4.3.1	MSA, phylogenetic tree calculations, and pairwise sequence identity	33
2.4.3.2	PPI network analysis	34
2.4.3.3	Functional annotation: functional domains prediction, motif analysis and homology modeling	35
2.5	Results and Discussion	38
2.5.1	Phase I	38
2.5.1.1	Sequence retrieval	38
2.5.1.2	Sequence homology search: Distribution and identification of HP homologs	38
2.5.2	Phase II	39
2.5.2.1	HPs in <i>V. corneae</i> are mostly involved in regulatory processes and the transportation of molecules	39
2.5.2.2	Binding proteins	47
2.5.2.3	Enzymes	47
2.5.2.4	Transport proteins	48
2.5.2.5	Structural proteins	50
2.5.2.6	Cellular/regulatory proteins	51
2.5.3	Transmembrane localization and signal peptide prediction	51
2.5.4	Physicochemical characterization of HPs	52
2.5.5	KEGG and Gene Ontology analysis	54

2.5.6	Characterization of Representative Proteins	54
2.5.6.1	Case study I: VICG00012	55
2.5.6.1.1	Identification and analysis of conserved motifs . .	55
2.5.6.1.2	Homology modeling	57
2.5.6.1.3	MSA and conserved domain analysis	57
2.5.6.1.4	Phylogenetic tree analysis and pairwise sequence identity calculations	57
2.5.6.1.5	PPI network analysis	58
2.5.6.2	Case study II: VICG01314	60
2.5.6.2.1	Identification and analysis of conserved motifs . .	60
2.5.6.2.2	Homology modeling	62
2.5.6.2.3	MSA and conserved domain analysis	62
2.5.6.2.4	Phylogenetic tree analysis and pairwise sequence identity calculations	62
2.5.6.2.5	PPI network analysis	62
2.5.6.3	Case study III: VICG01349	64
2.5.6.3.1	Identification and analysis of conserved motifs . .	64
2.5.6.3.2	Homology modeling	66
2.5.6.3.3	MSA and conserved domain analysis	66
2.5.6.3.4	Phylogenetic tree analysis and pairwise sequence identity calculations	66
2.5.6.3.5	PPI network analysis	66
2.5.6.4	Case study IV: VICG01687	68
2.5.6.4.1	Identification and analysis of conserved motifs . .	68
2.5.6.4.2	Homology modeling	70
2.5.6.4.3	MSA and conserved domain analysis	70
2.5.6.4.4	Phylogenetic tree analysis and pairwise sequence identity calculations	70
2.5.6.4.5	PPI network analysis	70
2.5.6.5	Case study V: VICG01723	71
2.5.6.5.1	Identification and analysis of conserved motifs . .	71
2.5.6.5.2	Homology modeling	74

2.5.6.5.3	MSA and conserved domain analysis	74
2.5.6.5.4	Phylogenetic tree analysis and pairwise sequence identity calculations	74
2.5.6.5.5	PPI network analysis	75
2.6	Conclusion	77
3	De novo Genome Assembly of <i>Microsporidia</i> sp. MB isolated from <i>Anopheles arabiensis</i> in Kenya	78
3.1	Chapter Overview	78
3.2	Keywords	79
3.3	Introduction	79
3.3.1	Microsporidia genome variations in structure are dictated by their mode of transmission	81
3.3.2	Current advances in Next-generation sequencing (NGS) tech- nologies and their application in microsporidia genomics studies	81
3.3.3	Overview of gene prediction tools	83
3.4	Methodology	84
3.4.1	Sample collection and DNA extraction	85
3.4.2	DNA sequencing and de novo assembly	85
3.4.3	Quality control and visualization of the assembled genome	86
3.4.4	<i>Microsporidia</i> sp. MB repeats and transposable elements annotation	87
3.4.5	Comparative phylogenomic analysis of <i>Microsporidia</i> sp. <i>MB</i> and the <i>Enterocytozoonida</i> group	87
3.5	Results and Discussion	88
3.5.1	Quality control analysis during and after sequencing . . .	88
3.5.1.1	Quality control of raw reads	88
3.5.2	<i>Microsporidia</i> sp. MB has a typical small genome	91
3.5.2.1	Visualization of the <i>Microsporidia</i> sp. MB genome architecture using BlobToolKit	92

3.5.2.2	Taxonomic classification of the assembled contigs reveal close homology to <i>V. corneae</i> and <i>N. ceranae</i> microsporidia species.	95
3.5.2.3	The small <i>Microsporidia</i> sp. MB genome could be attributed to the distribution of minimal repeats and transposable elements	99
3.5.2.4	The gene-dense <i>Microsporidia</i> sp. MB genome	102
3.5.3	Comparative phylogenomic analysis of <i>Microsporidia</i> sp. MB and the Enterocytozoonida group places the symbiont closest to <i>V. corneae</i>	104
3.6	Conclusion	106

4	<i>Microsporidia</i> sp. MB Genome Annotation and Functional Analysis	108
4.1	Chapter Overview	108
4.2	Keywords	109
4.3	Introduction	109
4.3.1	Microsporidia genomes: models of extensive gene loss driven by their obligate intracellular nature	109
4.3.2	Multifaceted computational annotation reveals important protein functions	109
4.3.3	Decay of components of the glycolytic pathway in microsporidia	110
4.3.4	The RNA interference (RNAi) machinery	110
4.3.5	Meiotic related genes	112
4.3.6	The Iron-Sulfur (Fe-S) cluster assembly machinery	112
4.4	Methodology	113
4.4.1	Overview of the annotation process	113
4.4.2	Protein classification	113
4.4.3	Gene Ontology (GO) analysis	114
4.4.4	KEGG Metabolic pathway analysis	114
4.4.5	Transmembrane localization and signal peptide predictions	114
4.4.6	Physicochemical properties	114
4.4.7	Conserved domains, motif analysis and homology modeling	114

4.4.8	Annotation of LRR proteins	115
4.4.9	Characterization of the glycolytic pathway in <i>Microsporidia</i> sp. MB	115
4.4.10	Characterization of Meiotic-related genes (MRGs) orthologs in <i>Microsporidia</i> sp. MB	116
4.4.11	Annotation of the RNAi mechanism in <i>Microsporidia</i> sp. MB	117
4.4.12	Phylogenetic analysis of Iron-Sulfur cluster assembly pro- teins in <i>Microsporidia</i> sp. MB	118
4.5	Results and Discussion	118
4.5.1	Proteins involved in enzymatic processes within the cell form the most significant proportion of the genome	118
4.5.1.1	Hydrolases and transferases are the most abun- dant enzyme classes in the <i>Microsporidia</i> sp. MB genome	119
4.5.1.2	Distribution of different enzyme sub-classes . . .	120
4.5.2	<i>Microsporidia</i> sp. MB contains multiple Leucine-rich repeat (LRR) proteins	123
4.5.2.1	Distribution of domains and repeats in <i>Microsporidia</i> sp. MB	123
4.5.2.2	Characterization of LRR proteins in <i>Microsporidia</i> sp. MB	124
4.5.3	<i>Microsporidia</i> sp. MB proteins are mainly involved in metabolic processes	127
4.5.3.1	GO analysis	127
4.5.3.2	KEGG pathway analysis	128
4.5.4	<i>Microsporidia</i> sp. MB has retained most of the genes in- volved in the glycolytic pathway	130
4.5.5	<i>Microsporidia</i> sp. MB contains all three key components of the RNAi machinery	131
4.5.5.1	Dicer	132
4.5.5.2	Argonaute	134
4.5.5.3	RNA-dependent RNA polymerase	136

4.5.6	The presence of several MRGs in <i>Mirosporidia</i> sp. MB suggests that it may occasionally undergo meiosis during sporogony	138
4.5.6.1	Role of the SMC proteins in meiosis	140
4.5.6.1.1	SMC1 [gene_603]	140
4.5.6.1.2	SMC2 [gene_89]	141
4.5.6.1.3	SMC3 [gene_1347]	143
4.5.6.1.4	SMC4 [gene_1051]	145
4.5.6.1.5	SMC6 [gene_941]	147
4.5.6.2	Sister chromatid cohesion 1 protein 3 (Scc3) [gene_1754]	149
4.5.6.3	DNA repair and recombination proteins in meiosis	151
4.5.6.3.1	RAD16 [gene_1760]	151
4.5.6.3.2	RAD51 [gene_1069]	153
4.5.6.3.3	RAD52 [gene_421]	155
4.5.6.3.4	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1/SNF2/RAD54 [gene_1398]	156
4.5.6.4	DNA mismatch repair (MMR) proteins in meiosis	158
4.5.6.4.1	MutL/MLH1 [gene_478]	158
4.5.6.4.2	Msh2/MutS [gene_1101]	160
4.5.6.4.3	MutS/MSH5 [gene_684]	162
4.5.6.4.4	PMS2/MutL [gene_1355]	164
4.5.6.4.5	MSH2/6 [gene_114]	166
4.5.6.5	MND1 family protein [gene_1920]	168
4.5.6.6	Sporulation protein RMD1 [gene_1498]	170
4.5.7	Phylogenetic Analysis of Iron-Sulfur Cluster Assembly Proteins in <i>Microsporidia</i> sp. MB	172
4.6	Conclusion	178
5	Importance of Research, Future Directions, and Concluding Remarks	179

5.1	Chapter 2: Structural and Functional Annotation of Hypothetical Proteins from <i>Vittaforma corneae</i> ATCC 50505 Using in silico Approaches	180
5.1.1	Importance of Research	180
5.1.2	Summary of Novel Findings	180
5.1.3	Further Implications and Future Directions	181
5.1.4	Concluding Remarks	181
5.2	Chapter 3: De novo Genome Assembly of <i>Microsporidia</i> sp. MB isolated from <i>Anopheles arabiensis</i> in Kenya	182
5.2.1	Importance of Research	182
5.2.2	Summary of Novel Findings	182
5.2.3	Further Implications and Future Directions	183
5.2.4	Concluding Remarks	183
5.3	Chapter 4: <i>Microsporidia</i> sp. MB Genome Annotation and Functional Analysis	184
5.3.1	Importance of Research	184
5.3.2	Summary of Novel Findings	184
5.3.3	Further Implications and Future Directions	184
5.3.4	Concluding Remarks	185
	List of References	186
	Appendices	233

List of Tables

1		xxix
2	List of Tools	xxxii
1.1	Overview of Mosquito-infecting Microsporidia as highlighted in [Andreadis, 2007].....	11
2.1	A summary of the selected templates used to model each HP . . .	37
2.2	A summary of the functional categories of the annotated HPs. . .	43
3.1	Distribution of classified raw reads before and after decontamination using Kraken2	92
3.2	Comparison of <i>Microsporidia</i> sp. MB with its closest relative <i>V. corneae</i> ATCC 50505	93
3.3	Distribution of different repeat types identified in the <i>Microsporidia</i> sp. MB assembly	100
4.1	An overview of the different stages of meiosis and the proteins involved in each stage.	139
4.2	An overview of <i>Microsporidia</i> sp. MB genes linked to the Fe-S cluster assembly machinery	173

List of Figures

1.1	Overview of the taxonomic classification of microsporidia [Bojko et al., 2022], figure redistributed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). Previous clade numbers are summarized at each node and to the right of the tree for ' <i>Glugeida+</i> ' (an informal label for <i>Hamiltosporidium</i> , <i>Neoflabelliforma</i> , <i>Areospora</i> , and <i>Astathelohania</i>).	4
1.2	Schematic representation of the microsporidia lifecycle (reproduced from [Bukhari et al., 2022] redistributed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)). The spore (A) injects the sporoplasm (C) into the cell via the polar tube (B). Microsporidia go through merogony (D-F) and sporogony (G) within the cell, resulting in more than one spore type (Hi, Hii) . .	6
2.1	A summary of the computational workflow designed in this study to annotate <i>V. corneae</i> HPs. The pipeline highlights two major phases: PHASE I. Sequence retrieval, similarity search, and test for homology. PHASE II. Diverse functional annotation using different tools to predict protein families, localization, motifs and domains identification, chemical properties of the HPs, sequence identity calculations, homology modeling, and phylogenetic analysis. . . .	28

2.2	A schematic representation of the reverse-blast method of testing true homology. To confirm the true homology of the HPs, the top 10 BLASTp hits were selected for each protein. The identification of true homologs was based on the appearance of the HP in the top three hits with high coverage and percent identity from the sequences analyzed.	32
2.3	(a) A pie-chart representation of the different protein family classes identified in <i>V. corneae</i> HPsA total of 162 HPs classified into respective functional groups based on functional annotations using several online tools. The number of HPs in each group is illustrated by the number following each group name and their respective percentage in brackets. The chart shows that the bulk of the proteins (37%) are involved in binding to different molecules (nucleic acids, proteins, ATP), followed by enzymes (23%), structural proteins (9%), regulatory proteins (8%), cell-cycle proteins (6%) and transport proteins (5%). Cellular and regulatory proteins include translation, transcription, and replication. The unknown proteins highlighted include those that did not have significant hits across the tools used. (b) A bar graph representation of the proportion of binding proteins. Nucleic acid-binding proteins, ATP, and protein-binding proteins. The numbers next to the labels indicate the total number of HPs in each group out of 60 binding proteins, with percentages. The order of each section is sorted by size, with the largest proportion of the binding HPs binding to DNA (13.22%) followed by metal ions (12.2%).	41

2.4	Overview of the physicochemical properties of the annotated HPs: (a) The GRAVY scores of less than 0 indicate the prediction of hydrophilic (globular) proteins, while scores larger than 0 are predicted to be hydrophobic (membrane) proteins. A total of 134 (81.7%) HPs were classified as hydrophilic, while 30 (18.3%) HPs were identified as hydrophobic. (b) Isoelectric point vs. molecular weight of HPs using the ProtParam tool. In red are the eccentric proteins with acidic pI (3–5) ($n = 20$; 12.2%) or basic pI (9–12) ($n = 49$; 29.87%), while those in black are proteins with pI range 5–9 ($n = 95$; 57.93%). The majority of HPs fall in this category	53
2.5	Functional and structural annotation of VICG00012. (a) The motif analysis heatmap generated from MEME Suite shows six highly conserved motifs across the PSA homologs, and motif 9 was missing in certain species. (b) Motif mapping to the homology model of the HP. Motifs numbers are preserved based on MEME outputs.(c) Motifs were mapped onto the MSA results of VICG00012 and its homologs, highlighting in purple all motifs conserved throughout the dataset. The conserved domain identified in this sequence belonged to the PRE1 superfamily.(d) Sequence identity heatmap highlighting a pairwise comparison of each of the sequences used in the phylogenetic analysis. A darker color shows a closer identity to respective sequences. Two distinct clusters are observed with higher identity in the <i>Encephalitozoon</i> family compared to the clade containing the HP of interest.	56

2.6	A PPI network analysis using the STRING database, with the query HPs highlighted in red (L2GQV7). Each node represents a protein, and a total of 10 interacting proteins (nodes) were observed, with 55 edges indicating both physical and functional protein associations. The color of the edges indicates the type of interaction evidence, with pink edges indicating experimentally-determined associations and purple edges indicating protein homology. The analysis revealed that the HP has close homology to a proteasome alpha 1 domain-containing protein (L2GLK0) and is experimentally linked to similar proteins. The protein-protein interaction (PPI) enrichment p-value is 4.66×10^{-15} , indicating that the clustered proteins in this group are at least partially biologically connected.	59
2.7	Structural and functional analysis of VICG01314. (a) Motif analysis of the HP showing nine conserved regions. (b) Mapping of the motifs onto the 3D structure of the HP (c) MSA illustrating, in purple, several conserved regions in this set of sequences. (d)Phylogenetic analysis of VICG01314 using the Maximum Likelihood method shows that the HP VICG01314 clustered away from the rest of the clades but was closer relative to <i>H. eriocheir</i> and <i>E. hepatopenaei</i> (The heatmap scale ranges from 0-100%).	61
2.8	Protein association prediction shows VICG01314 to be in close proximity to other regulatory and enzymatic proteins with varying functions such as translation initiation factor, synthases, kinases, and GTP-binding proteins	63

2.9	Structural and functional analysis of VICG01349.(a) Motif heatmap of VICG1349 showing conservation of 7 motifs across all species. Species and protein IDs are indicated in the legend.(b) The modeled 3D structure of HP VICG01349 with the seven conserved motifs mapped onto the structure. (c) Multiple sequence alignment and motif mapping on VICG01349. Two main domains (the winged-helix DNA-binding domain and the Dimerization domain) are also indicated spanning the region across the sequence as a diagrammatic representation. (d) Tree calculations using the Maximum Likelihood method and sequence identity heatmap of VICG01349 and its homologs illustrate the close relationship between proteins within the same clade (heatmap scale from 0-100%).	65
2.10	PPI analysis of HP VICG01349 using STRING analysis. Close associations with proteins involved in cell cycle processed is indicated.	67
2.11	(a) Motif analysis heatmap of the protein VICG01687, highlighted in red. A large number of motifs were conserved among these proteins. (b) The 3D model of VICG01687 as predicted using Modeller. Motif mapping is based on the most conserved motif identified using MEME. (c) Multiple sequence alignment and motif mapping of VICG01687. Two main domains are mapped onto the diagrammatic representation of the one-dimensional sequence illustrated below the alignment. (d) Phylogenetic analysis using the Maximum Likelihood method and pairwise identity calculations of VICG01687 and its homologs. Two distinct clusters are observed (Heatmap scale ranging from 0% to 100%).	69
2.12	STRING analysis of the proteins interacting with VICG01687. The query node indicates the HP VICG01687. Its closest neighbors are proteins from the helicase and argonaute family.	71

2.13	(a) Motif analysis heatmap of VICG01723 shows a high number of conserved motifs. (b) The 3D modeled structure of VICG01723 with motif mapping. (c) Multiple sequence alignment of the HP VICG01723 with motif mapping respective to MEME results. Motif mapping is based on the numbering of the identified motifs as determined using MEME. MEME analysis was used to detect conserved sequence motifs across the protein family, which are represented in the alignment. An N-terminal domain (DEAD-BOX HELICASE) and a C-terminal domain are highlighted in the sequence.(d) phylogenetic classification of HP VICG01723 using the Maximum Likelihood tree calculation method and pairwise sequence calculations highlight the sequence identity values among the homologous species (Heatmap scale: 0-100%).	73
2.14	PPI network analysis of VICG01723 (query node in red) showing close association with ribosomal proteins.	76
3.1	(a) The proportion of unique reads in each dataset based on FastQC analysis. (b) Number of total unique reads retained post-decontamination and host removal.	90
3.2	Superkingdom taxonomic classification of raw reads before and after decontamination using Kraken2. Only unclassified and eukaryotic reads were used for de novo assembly	91

3.3	<i>Microsporidia</i> sp. MB genome structure. (a) The snail plot created using BlobToolKit highlights the genome metrics. A basic legend of the assembly statistics is provided in the top left corner. The outer blue rings highlight the GC content over the assembly, corresponding to each of the contigs in the grey ring. The overall GC content of the assembly is summarized at the bottom left corner. The light grey rings show the total length of the assembly. The red line highlights the longest scaffold, while the dark and light orange rings show the length of the contigs at which 50% and 90% of the total assembly are represented, respectively. (b) A pie-chart representation of the BUSCO gene completeness analysis demonstrates a total of 81% single-copy core microsporidia genes present in <i>Microsporidia</i> sp. MB.	94
3.4	Species distribution of top BLAST hits	96
3.5	A homology search using SprayNPray v3 [Garber et al., 2022] identified all cleaned contigs were microsporidia. The most abundant species homologous to the <i>Microsporidia</i> sp. MB contigs were <i>N. ceranae</i> and <i>V. corneae</i>	98
3.6	Distribution of different repeat types in <i>Microsporidia</i> sp. MB. The presence of short tandem repeats in small proportions was identified in this dataset.	101
3.7	A representation of the genome structure of <i>Microsporidia</i> sp. MB, displaying contigs larger than 10Kb, highlighting gene density and organization per sequence. The identified genes correspond to the BUSCO identification of GeneMark-ES annotated proteins. A color-coded legend is included, categorizing the general function of each putative protein into 12 distinct categories. The plot was created in RStudio using the gggenes package (https://cran.r-project.org/web/packages/gggenes).	103

3.8	Phylogenomic analysis of common BUSCO orthogroups found in Enterocytozoonida (left) and their respective genome assembly completeness using BUSCO (right). <i>Microsporidia</i> sp. MB is in bold. <i>Nosema granulosis</i> was used as an outgroup. Common orthogroups were aligned using MUSCLE [Edgar et al., 2004] and trimAl [Capella-Gutiérrez et al., 2009]. The maximum likelihood tree (1,000 bootstraps) was constructed using IQ-TREE v1.6.12 [Nguyen et al., 2016] based on the 92 single-copy orthologous genes (with the best model selection (LG+F+I+G4) from ModelFinder [Kalyaanamoorthy et al., 2017]. Bootstrap percentage values are shown on each node, and the scale bar at the bottom of the tree indicates total changes per site.	105
4.1	A schematic representation of the RNAi machinery highlighting the steps where the Dicer, Argonaute and RNA-dependent RNA polymerase have active roles in.	111
4.2	A pie-chart representation of the proportion of different protein family categories in <i>Microsporidia</i> sp. MB using InterProScan. Seven main protein superfamily groups were identified, with most of the proteins classified as enzymes.	119
4.3	The distribution of primary enzyme groups using InterProScan, highlighting hydrolases and transferases as the most abundant enzyme classes in <i>Microsporidia</i> sp. MB. These two enzymes function in energy metabolism and transport of nutrients respectively. . . .	120
4.4	Distribution of different enzyme classes and sub-classes in <i>Microsporidia</i> sp. MB	121

4.5	Sequence distribution of the different domains and repeats predicted using InterProScan. (a) A bar plot of the key domains isolated from the genome showing a large set of domains involved mainly in enzymatic activities. A zoomed image of the main domains excluding the outlier (where less than 10 sequences were represented in each domain) is shown on the right panel. (b) The distribution of repeat regions is depicted in this barplot. Leucine-rich repeat regions (LRR) and WD40 were the most abundant. . .	124
4.6	Identification and analysis of leucine-rich repeats (LRR) motifs across 9 contigs. LRR motifs 1 (red) and 2 (blue) were identified in the nine contigs using MEME Suite. Moreover, prediction of signal peptides using TargetP v2.0 and SignalP v6.0 (green and brown respectively) in 2 of these sequences (NODE_405 and NODE_1665).	126
4.7	A 3D model of gene_679 (NODE_405) with motif mapping. The structure was modeled on PRIMO. The identified LRR motifs are numbered based on the MEME output and color represented is consistent with Figure 4.6 above.	127
4.8	A highlight of the protein sequence distribution in each primary GO term (Biological Process, Cellular Component and Molecular Function) shows that a bulk of the putative proteins are involved in metabolic processes, localized within the intracellular anatomical structure (nucleus and cytoplasm) and involved in enzymatic activities.	128
4.9	KEGG metabolic pathway analysis. A barplot representation of the main KEGG categories shows most of the pathways identified within the genome are involved in organismal systems and metabolism.	129
4.10	Sequence distribution in each KEGG pathway highlighting the top 25 pathways represented in the dataset. A majority of the sequences were involved in purine and thiamine metabolism, cellular processes and genetic information processing.	130

4.11	Overview of the loss and retention of key genes involved in the glycolytic pathway genes across different microsporidia species. Retained genes are highlighted in light blue in the heatmap, while those that have been lost are highlighted in gray. <i>Microsporidia</i> MB has retained most of the genes involved in glycolysis except for Pyruvate kinase, similar to <i>V. corneae</i> and <i>N. granulosis</i> . Comparatively, <i>E. aedis</i> , <i>E. cuniculi</i> and <i>E. hepatopenaei</i> appear to have lost all these genes, suggesting a total reliance on the host's glycolysis function. The maximum likelihood phylogenomic tree on the left was generated using IQ-TREE v1.6.12 based on the 196 common BUSCO single-copy orthologous genes found in all species. Percent bootstrap support values are shown on each node and the branch length scale bare at the bottom of the tree.	131
4.12	Analysis of the Dicer ortholog in <i>Microsporidia</i> sp. MB (a) A rooted phylogenetic tree of Dicer genes in various microsporidian species. The tree was constructed using The scale for the branch length is shown at the bottom. (b) A schematic representation of the position of conserved domains in the multiple sequence alignment is shown. An insert was observed between the RIBOc and Rnc domains in 4 of the species including <i>Microsporidia</i> sp. MB, <i>E. breve</i> , <i>P. epiphaga</i> and <i>P.philotis</i>	133
4.13	Analysis of Argonaute orthologs. (a) A rooted phylogenetic tree of Argonaute orthologs in microsporidia. The tree was constructed on protein sequences of the Argonaute orthologs. A thousand bootstraps are indicated at the nodes. The scale for the branch length is shown at the bottom. (b) A schematic representation of the position of conserved Piwi-like and PAZ domains mapped onto the multiple sequence alignment.	135
4.14	Analysis of the RNA-dependent RNA polymerase. (a) Phylogeny studies indicate the similarity between this ortholog in <i>Microsporidia</i> sp. MB and other microsporidias. (b)Identification of the conserved RdRP domain across different species.	137

4.15	Overview of the SMC1 ortholog in <i>Microsporidia</i> sp. MB.	141
4.16	Analysis of the gene_89 SMC2 ortholog.	142
4.17	The SMC3 ortholog in <i>Microsporidia</i> sp. MB.	144
4.18	Analysis of the similarity between the SMC4 orthologs in <i>Microsporidia</i> sp. MB and other species.	146
4.19	A highlight of SMC6 orthologs in microsporidia.	148
4.20	Phylogenetic comparison and msa of Sister chromatid cohesion 1 protein 3.	150
4.21	The RAD16 DNA repair and recombination protein in microsporidia.	152
4.22	Analysis of the relationship between RAD51 orthologs in microsporidia.	154
4.23	Analysis of RAD52 in microsporidia.	155
4.24	Analysis of the SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1/SNF2/RAD54.	157
4.25	Tree calculations and sequence analysis of the MutL/MLH1 gene.	159
4.26	Tree calculations and sequence analysis of the Msh2/MutS gene orthologs in microsporidia	161
4.27	Tree calculations and sequence analysis of the MutS/MSH5 gene.	163
4.28	Tree calculations and sequence analysis of the PMS2/Mut gene.	165
4.29	Tree calculations and sequence analysis of the MSH2/6 gene.	167
4.30	Tree calculations and sequence analysis of the MND1 family protein.	169
4.31	Tree calculations and sequence analysis of the Sporulation protein RMD1.	171
4.32	Analysis of the mitochondrial ferredoxin homolog in <i>Microsporidia</i> sp. MB gene_19. (a) Tree calculations done using Neighbor-joining method shows high similarity within this group. Bootstrap support values are highlighted on each branch, and the scale is shown at the bottom of the tree. (b) Multiple sequence alignment of gene_19 and its mitochondrial ferredoxin homologs. Two inserts were iden- tified in <i>H. eriocheir</i>	174

4.33	Analysis of the cysteine desulfurase homolog in <i>Microsporidia</i> sp. MB gene_371. (a) Tree calculations using the neighbour-joining method show high similarity. Bootstrap support values are highlighted on each branch, and the scale is shown at the bottom of the tree. (b) Multiple sequence alignment of gene_371 and its homologs show high similarity.	175
4.34	Analysis of the molecular chaperone HscB homolog in <i>Microsporidia</i> sp. MB gene_1659. (a) The generated neighbour-joining trees show high similarity. Bootstrap support values are indicated on each branch. The scale is shown at the bottom of the tree. (b) Multiple sequence alignment of gene_1659, highlighting an insert in <i>Microsporidia</i> sp. MB (red box).	176
4.35	Analysis of the Fe-S cluster assembly enzyme IscU homolog in <i>Microsporidia</i> sp. MB gene_1773. (a) Tree calculations done with neighbour-joining method is illustrated here. Bootstrap support values are shown in each branch indicating high similarity. The scale is shown at the bottom of the tree. (b) Multiple sequence alignment of gene_1773 and other	177

List of Abbreviations

ICIPE	International Center for Insect Physiology and Ecology
RUBI	Reasearch Unit in Bioinformatics
CHPC	Centre for High Performance Computing
OWSD	Organization for Women in Science for the Developing World
BMGF	Bill & Melinda Gates Foundation
CIFF	Children's Investment Fund Foundation
DNBSeq	DNA Nanoball Sequencing
HP	Hypothetical Proteins
SSU rRNA	Small Sub-unit ribosomal RNA
ITS	Internal Transcribe Spacer
OGRI	overall genome relatedness indices
ANI	average nucleotide identity
PPI	protein-protein interaction
BLAST	Basic Local Alignment Search Tool
HMMs	hidden Markov models
SVM	support vector machines
PLMSearch	Protein language model search
ExPASy	Expert Protein Analysis System
pI	isoelectric point

GRAVY	grand average of hydropathicity
MEME	Multiple Expectation-maximization for Motif Elicitation
PFS	protein functional sites
PLI	protein-ligand interactions
fDPD	function- and interaction-annotated protein domain profile database
PSSMs	position-specific matrices
SMART	Simple Modular Architecture Research Tool
SP	Signal Peptide
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
DIP	Database of Interacting Proteins
CATH	protein structure classification
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
KEGG	Kyoto Encyclopedia of Genes and Genomes
KofamKOALA	KEGG Orthology And Links Annotation
MUSCLE	Multiple Sequence Comparison by Log-Expectation
MEGA	Molecular Evolutionary Genetics Analysis
CDD	Conserved Domain Database
GO	Gene Ontology
PRIMO	Protein Interactive Modeling
DHFR	dihydrofolate reductase
ABC	ATP-Binding Cassettes
MSA	Multiple sequence alignment
3D	three-dimensional
Fe-S	Iron-Sulfur

MICSWaP	microsporidian spore wall superfamily protein
PSA	proteasome alpha subunit
NGS	Next-generation sequencing
BGI	Beijing Genomics Institute
cPAS	combinatorial probe-anchor synthesis
BWA	Burrows-Wheeler Aligner
QUAST	QUality Assessment Tool
BUSCO	Benchmarking Universal Single-Copy Ortholog
SINEs	short interspersed nuclear elements
LINEs	long interspersed nuclear elements
LTR	long tandem repeat element
STR	short tandem repeats
RNAi	RNA interference
RISC	RNAi-induced silencing complex
DSBs	DNA doublestrand breaks
LRR	Leucine-rich repeat
SCC	Sister chromatid cohesion

List of Tools

Table 2: [A list of bioinformatics tools and resources used.]

Annotation Type	Software	URL	Reference
<i>Sequence retrieval and similarity search</i>	NCBI-BLAST	https://blast.ncbi.nlm.nih.gov/B	[Johnson et al., 2008]
	MicrosporidiaDB	https://microsporidiadb.org/	[Aurrecoechea et al., 2011]
	UniProt	https://www.uniprot.org/	[Bateman, 2019, Bateman et al., 2021]
	SWISS-MODEL	https://swissmodel.expasy.org	[Waterhouse et al., 2018]

Continued on next page

Table 2: [A list of bioinformatics tools and resources used.] (Continued)

Annotation Type	Software	URL	Reference
<i>Programming languages for script generation</i>	Python v. 3.8	https://www.python.org/	[Rossum, 2017]
	Jupyter Notebook	https://jupyter.org/	[Silaparasetty, 2020]
<i>Protein family classification</i>	Pfam	http://pfam.xfam.org	[Mistry et al., 2021]
	InterPro	https://www.ebi.ac.uk/interpro/	[Blum et al., 2021]
	HMMer	https://www.ebi.ac.uk/Tools/hm	[Potter et al., 2018]
	PANTHER	http://www.pantherdb.org	[Mi et al., 2021]
	PRINTS	http://130.88.97.239/PRINTS/ir	[Attwood, 2002]
	PIRSF	https://proteininformationresour	[Wu et al., 2004]
<i>Sequence alignment and phylogenetic analysis</i>	MEGA X	https://www.megasoftware.net/	[Kumar et al., 2012, 2018]
	MUSCLE	NA	[Edgar et al., 2004]
	SWISS-MODEL	https://swissmodel.expasy.org	[Waterhouse et al., 2018]

Continued on next page

Table 2: [A list of bioinformatics tools and resources used.] (Continued)

Annotation Type	Software	URL	Reference
	Dendroscope v. 3.7.4	https://uni-tuebingen.de/en/fakultaeten/mathematische_naturwissenschaftliche_fakultaet/fachbereiche/informatik/lehrstuehle/bioinformatik/software/dendroscope/	[Huson et al., 2007]
<i>Motif analysis and domain functional annotation</i>	MEME Suite (MEME & MAST)	https://meme-suite.org	[Bailey et al., 2015]
	EXPASY-ScanProsite	https://prosite.expasy.org/scanprosite/	[de Castro et al., 2006]
	CDD Search	https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.htm	[Lu and Salzberg, 2020]
	SMART	https://smart.embl.de	[Letunic et al., 2021]
<i>Characterization of physico-chemical properties</i>	EXPASY-PROTPARAM	https://web.expasy.org/protparam/	[Gasteiger et al., 2005b]
	SWISS-MODEL	https://swissmodel.expasy.org	[Waterhouse et al., 2018]

Continued on next page

Table 2: [A list of bioinformatics tools and resources used.] (Continued)

Annotation Type	Software	URL	Reference
<i>Subcellular localization</i>	TMHMM v. 2.0	http://www.cbs.dtu.dk/services/	[Sonnhammer and Krogh, 2008]
	TargetP v. 2.0	http://www.cbs.dtu.dk/services/	[Armenteros et al., 2019]
	SignalP v. 5.0	http://www.cbs.dtu.dk/services/	[Almagro Armenteros et al., 2019]
	Phobius	https://www.ebi.ac.uk/Tools/pfa	[Käll et al., 2004]
<i>Protein-protein association network</i>	STRING v. 11 (Search tool for the retrieval of interacting genes and proteins)	https://string-db.org/	[Szklarczyk et al., 2021]
<i>Gene Ontology</i>	KEGG	https://www.genome.jp/kegg/	[Kanehisa et al., 2016]
<i>Homology Modeling</i>	PRIMO	https://primo-oligo.rubi.ru.ac.za/	[Hatherley et al., 2016]
	MODELLER	https://salilab.org/modeller/	[Söding et al., 2005]
	PROSA-web	https://prosa.services.came.sbg.a	[Wiederstein and Sippl, 2007]
	SWISS-MODEL	https://swissmodel.expasy.org	[Waterhouse et al., 2018]

Continued on next page

Table 2: [A list of bioinformatics tools and resources used.] (Continued)

Annotation Type	Software	URL	Reference
	VERIFY3D	https://saves.mbi.ucla.edu/	[Luthy et al., 1997]
	PROCHECK	https://www.ebi.ac.uk/thornton-srv/software/PROCHECK/	[Pontius et al., 1996]
	QMEAN	http://swissmodel.expasy.org/qm	[Benkert et al., 2009]
	SWISS-MODEL	https://swissmodel.expasy.org	[Waterhouse et al., 2018]

Research Outputs

Primary Publications

1. **Ang'ang'o, L.M.**, Herren, J.K., Tasthan Bishop, Ö. (2023) Structural and Functional Annotation of Hypothetical Proteins from the Microsporidia Species *Vittaforma corneae* ATCC 50505 Using in silico Approaches. *Int. J. Mol. Sci.* 2023, 24, 1–27, doi:10.3390/ijms24043507.
2. **Ang'ang'o, L. M.**, Waweru, J. W., Makhulu, E. E., Wairimu, A., Otieno, F. G., Onchuru, T., Tasthan Bishop, Ö., Herren, J. K. (2024). Draft genome of *Microsporidia* sp. MB—a malaria-blocking microsporidian symbiont of the *Anopheles arabiensis*. *Microbiology Resource Announcements*, 13(4), e00903-23, doi:10.1128/MRA.00903-23.

Conference Presentations (Oral)

1. ANTI-VEC 4th Annual Meeting, Kilifi, Kenya.
2. 16th International Workshop on Opportunistic Protists (IWOP), Lausanne, Switzerland.

Thesis Overview

The persistent threat of malaria, transmitted by the notorious *Anopheles* mosquito, continues to pose a significant public health challenge in many parts of the world. Innovative strategies are therefore urgently needed to complement existing control measures. One promising approach involves the use of symbiotic microorganisms, such as microsporidia, that can inhibit the development of malaria parasites within their mosquito hosts. Microsporidia are a diverse group of obligate intracellular fungi-like organisms infecting a wide host range. These ubiquitous unicellular eukaryotes have a significant impact on human health, agriculture and the environment. *Microsporidia* sp. MB, denoted as "MB" to taxonomically represent a set of unclassified/unidentified Microsporidia within this genus, has been identified as a malaria-blocking microsporidia associated with *Anopheles* mosquitoes and represents a unique example of symbiosis within their hosts. Through its intricate interactions with both the mosquito vector and the malaria parasite, *Microsporidia* sp. MB has demonstrated remarkable potential as a biocontrol agent against malaria transmission. However, the biology and genomic landscape of this microorganism remain relatively understudied.

This thesis focused on understanding the biology of the malaria-blocking *Microsporidia* sp. MB through bioinformatics analysis and aimed to assemble and annotate its genome. Unlocking the genetic blueprint of *Microsporidia* sp. MB is a critical step toward harnessing its potential in the global fight against malaria. By revealing the genes and molecular pathways that enable *Microsporidia* sp. MB to block malaria parasite development, genome sequencing promises to provide insights into host-microbe interactions, potential impacts on mosquito behavior, and population dynamics. Additionally, sequencing this genome can provide valuable information on its evolutionary history and relationship to other microsporidia,

helping in further understanding how these organisms have adapted to their hosts and developed mechanisms to block pathogens.

This document is organized into five research chapters, with chapters 2 - 4 highlighting the main approaches used to reach the main objectives of this research work.

Chapter 1 provides an extensive literary overview of microsporidia, highlighting their obligate intracellular nature, main biological characteristics, taxonomic classification, and complex host-microbe interactions. This chapter also highlights the research gap and main objectives of the study.

Chapter 2 describes the structural and functional annotation of hypothetical proteins from the *Vittaforma corneae* ATCC 50505 genome. *V. corneae* ATCC 50505 was selected as a model for *Microsporidia* sp. MB due to their close phylogenetic relationship and potential similarities in genomic features. Moreover, *V. corneae* ATCC 50505 has been widely studied, providing a valuable reference for comparative genomics between the two species. This chapter highlights the design and successful use of an exhaustive computational pipeline to efficiently annotate these proteins using *V. corneae* as a model. This work has been published in [Ang'ang'o et al., 2023].

Chapter 3 focuses on the assembly of the *Microsporidia* sp. MB genome isolated from *Anopheles arabiensis* infected tissues using the Next-generation sequencing (NGS) platform DNA Nanoball Sequencing (DNBSeq) developed by the Beijing Genomics Institute (BGI). The preparation and sequencing of this microorganism and the analysis of its genome structure are carefully described here and has been reported in [Ang'ang'o et al., 2024].

Chapter 4 builds up from chapter 3 and annotates the genetic makeup of *Microsporidia* sp. MB. This chapter uses the pipeline modeled in chapter 2 to classify and characterize the function and structure of the putative genes found in *Microsporidia* sp. MB. This section also focuses on several key pathways and processes that are unique in microsporidia and investigates their existence in this symbiont. These include the glycolytic pathway, the RNA interference machinery and mode of sexual reproduction through meiosis. This analysis is described by identifying and annotating the genes involved in each process.

In **Chapter 5**, the main findings from Chapters 2–4 are comprehensively synthesized, presenting a cohesive summary of the core outcomes of the project while highlighting potential future research avenues. In this concluding chapter, the integrity of the project is demonstrated by summarizing the results from genome annotation, sequencing, and comparative analyses into a cohesive narrative that ties back to the research objectives. The pipeline developed and applied throughout the project, from protein annotation in *V. corneae* to the genome of *Microsporidia* sp. MB, showcases a consistent methodological approach, ensuring that the findings contribute meaningfully to both the understanding of microsporidian biology and potential malaria control efforts. This integrative approach reflects the scientific coherence of the project and highlights the alignment between its aims, methods, and outcomes.

A key part of the integrity of the project lies in the strong and consistent methods used throughout the research. The pipeline developed to analyze proteins in *V. corneae* was adapted for the genome of *Microsporidia* sp. MB, ensuring the same approach was applied across the study. This method provided a solid basis for gene analysis and allowed for useful comparisons between the two species. This careful design shows how each step of the project was aligned with its overall goals. The thesis demonstrates how the objectives, methods, and results come together to build a complete understanding of microsporidian genomics. By combining different data and techniques, the study addresses key research questions and contributes to the wider field of microsporidia genome annotation. The findings also offer insights into important genomic features and symbiotic relationships that could help develop new biocontrol strategies.

Chapter 1

Literature Overview

1.1 Background

Microsporidia are a group of eukaryotic spore-forming microorganisms considered the earliest-diverging and highly specialized fungi [Capella-Gutiérrez et al., 2012]. They infect a wide range of hosts, from vertebrates to invertebrates, including humans [Wittner and Weiss, 1999]. Due to their obligate intracellular nature, these organisms exhibit unique biological and life cycle adaptations that have diverse impacts on their hosts [Keeling and Slamovits, 2004, Keeling and Corradi, 2011]. They display considerable diversity in morphology and host specificity, with some being generalists and others showing a more restricted host range [Desjardins et al., 2015, Murareanu et al., 2021]. They are known for their unique reductive and derived features, including compact genomes, reduced metabolic capacity, and conservation of high proportions of genes involved in host-parasite interactions [Slamovits et al., 2004, Keeling and Corradi, 2011, Pombert et al., 2012, Peyre-taillade et al., 2011, Corradi, 2015, Pombert et al., 2015]. Microsporidia are often pathogenic in humans, primarily affecting immunocompromised individuals such as patients with HIV/AIDS and organ transplant recipients [Didier, 2005]. In these cases, microsporidia species have been linked to ocular, intestinal, pulmonary, and renal diseases [Wittner and Weiss, 1999]. Microsporidia also have a significant impact on agriculture and the economy. Invertebrates are the largest group of organisms infected with microsporidia. Most of these are pathogenic species affecting insect populations. Infections in silkworms, honeybees, and shrimp threaten the

agricultural sector [Higes et al., 2008, Meng et al., 2018, Stentiford et al., 2019, Wang et al., 2023]. In contrast, microsporidia infections in agricultural pests such as grasshoppers and corn borers have an advantageous impact on controlling these pest populations [Bjørnson and Oi, 2014]. Therefore, it is essential to understand the dynamics of infection between microsporidia and their host species to leverage the unique properties of these organisms for our benefit.

1.1.1 The complex taxonomic classification of microsporidia

To date, approximately 1600 species and over 220 genera of microsporidia have been identified [Bojko et al., 2022]. Classifying these complex organisms is continuously evolving as new species are discovered and older species are re-characterized [Park and Poulin, 2021, Bojko et al., 2022, Wang et al., 2023]. At present, multiple features are used to extensively describe and group microsporidian species into their respective clades. These include molecular characterization based on the highly conserved Small Sub-unit ribosomal RNA (SSU rRNA) and the Internal Transcribe Spacer (ITS) regions and ultrastructural analysis such as spore morphology, life cycle, and pathology [Vossbrinck et al., 2004, Tokarev et al., 2015]. Microsporidia are generally subdivided into terrestrial and aquatic species depending on the infected host environment. These classes are named Terresporidia, Aquasporidia, and Marinosporidia, representing terrestrial, freshwater, and marine habitats [Vossbrinck and Debrunner-Vossbrinck, 2005]. They were initially grouped into five clades using informal taxonomic numbers (Clades I–V) [Park and Poulin, 2021], but this has recently been updated to include formal names: Glugeida, Nosematida, Enterocytozoonida, Amblyosporida, Neopereziida, and Ovavesiculida respectively, based on the assessment of several genera of microsporidia [Bojko et al., 2022]. The analysis of rRNA sequences for 126 described genera has identified four previously recognized main clades but also highlighted the need for reevaluation within some clades [Bojko et al., 2022]. The current classification of microsporidia has evolved from the original morphology-based taxonomy to relying more on molecular characterization using distinct molecular markers, particularly SSU rRNA. This has led to the classification of clades based on these microorganisms’ evolutionary history and ecological niches. [Voss-

brinck and Debrunner-Vossbrinck, 2005, Park and Poulin, 2021, Bojko et al., 2022]
(Figure 1.1).

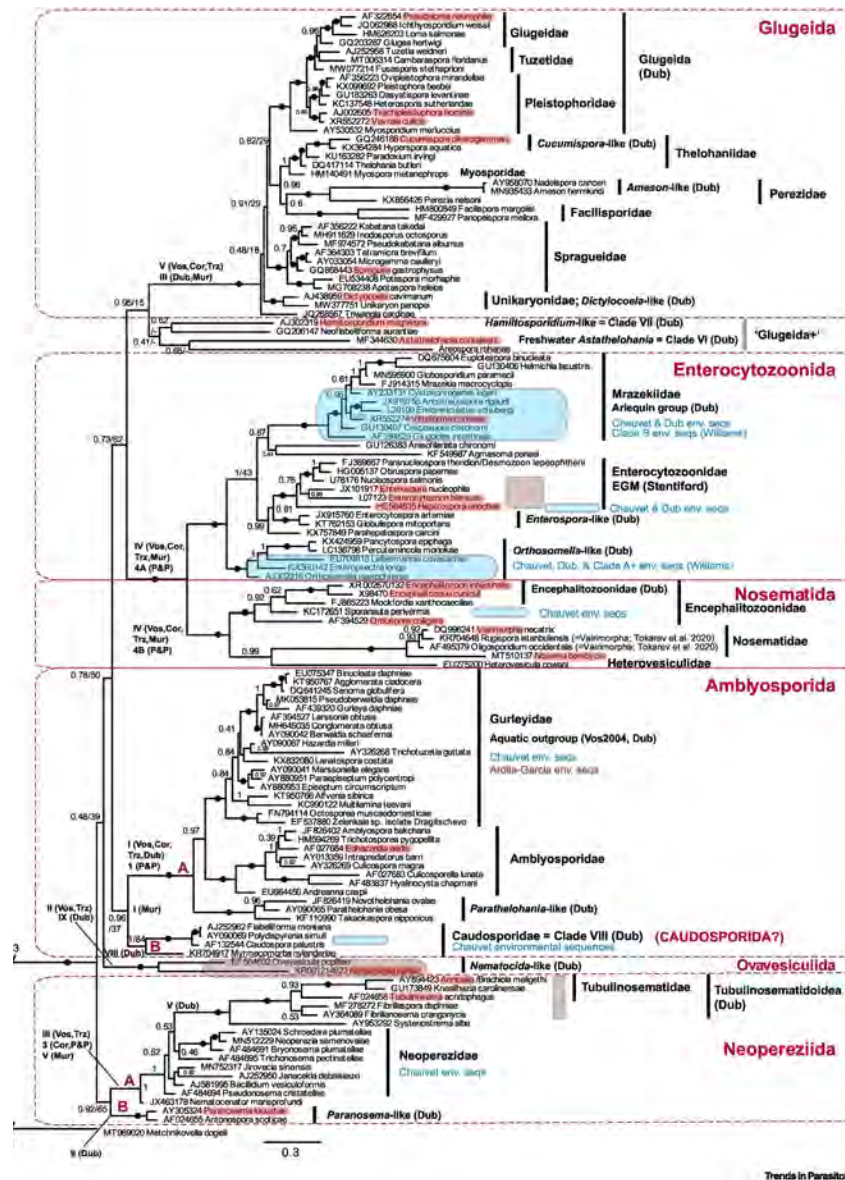


Figure 1.1: Overview of the taxonomic classification of microsporidia [Bojko et al., 2022], figure redistributed under a [Creative Commons Attribution 4.0 International License \(CC BY 4.0\)](#). Previous clade numbers are summarized at each node and to the right of the tree for 'Glugeida+' (an informal label for *Hamiltosporidium*, *Neoflabelliforma*, *Areospora*, and *Astatelohania*).

1.1.2 The microsporidia life cycle

The life cycle of microsporidia can be divided into three phases: the infective or environmental phase, the proliferative phase, and the sporogony or spore-forming phase [Cali and Takvorian, 2014]. In the infective phase, the resistant spores are released into the environment and can infect various hosts, including animals and protists [Willis and Reinke, 2022]. The spores have a rigid cell wall consisting of three layers: the inner endospore, the outer exospore, and the plasma membrane [Yang et al., 2018]. Once inside a compatible host, the spores germinate, and the polar tube, a specialized invasion organelle, is expelled to pierce the host cell membrane and transfer the spore contents into the host cell cytoplasm [Williams, 2009]. The parasite then undergoes several steps, including host cell invasion, proliferation within the host tissue, and spore formation [Willis and Reinke, 2022]. The proliferative phase involves the rapid multiplication of the parasite within the host cell, often causing hypertrophied cells and reorganizing host cell organelles and microtubules [Williams, 2009]. During the sporogony phase, the parasite forms mature spores, which can be released into the environment to continue the lifecycle [Cali and Takvorian, 2014] (Figure 1.2).

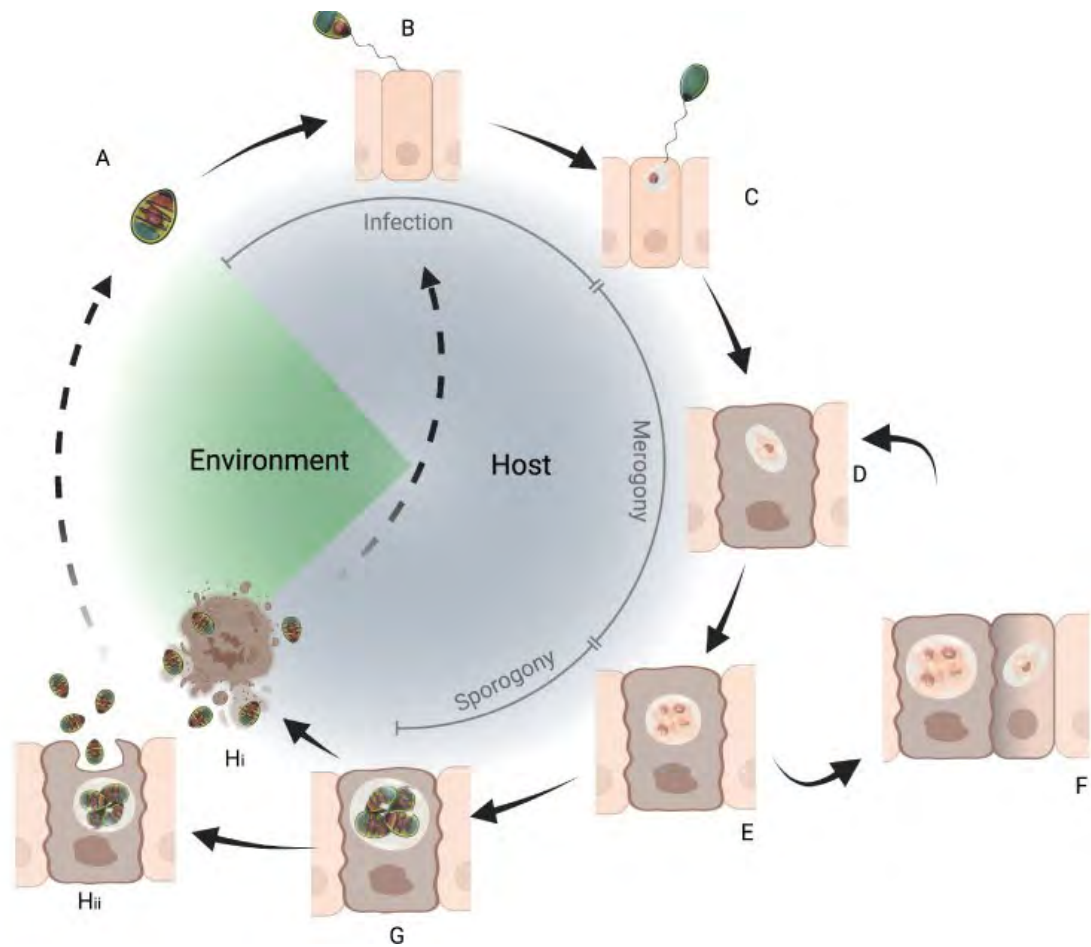


Figure 1.2: Schematic representation of the microsporidia lifecycle (reproduced from [Bukhari et al., 2022] redistributed under a [Creative Commons Attribution 4.0 International License \(CC BY 4.0\)](#)). The spore (A) injects the sporoplasm (C) into the cell via the polar tube (B). Microsporidia go through merogony (D-F) and sporogony (G) within the cell, resulting in more than one spore type (Hi, Hii)

The factors influencing microsporidia infection include temperature, seasonality, nutrient availability, the presence or absence of microbes, and the genetics of the host [Willis and Reinke, 2022]. Host genomic regions have been identified that influence resistance to microsporidia [Willis and Reinke, 2022]. Further research is needed to uncover the molecular mechanisms of microsporidia host specificity in more detail [Willis and Reinke, 2022].

1.1.3 Microsporidia genomes and genomic variations

The genomes of microsporidia have undergone massive reductions, particularly in nucleic and amino acid metabolism genes, as an adaptation to their obligate parasitic life cycle [Keeling et al., 2005, Huang et al., 2023]. Additionally, they are characterized by many transposable elements shaped by gene loss and horizontal gene transfers [Corradi, 2015, Pombert et al., 2012]. Genome variations in size and architecture in microsporidia are primarily due to non-coding material such as intergenic spaces and retrotransposons [Xu et al., 2006, Williams et al., 2008]. This has resulted in the reduction of their genome size and the loss of genes associated with various metabolic pathways, such as the tricarboxylic acid cycle [Katinka et al., 2001, Keeling et al., 2010, Luo et al., 2021], nucleotide, and amino acid synthesis [Heinz et al., 2014, Dean et al., 2016, Alexander et al., 2016, Melnikov et al., 2018]. The compacted microsporidia genome is considered the smallest among primary eukaryotes and even falls within the range of bacterial genome sizes. Microsporidia have experienced a high degree of organelle simplification, including the loss of canonical mitochondria [Germot et al., 1997, Burri et al., 2006, Waller et al., 2009]. The high levels of intraspecific genomic diversity observed in microsporidia genomes raise essential questions about the factors that drive this variation [Katinka et al., 2001, Cuomo et al., 2012] and are crucial for unraveling their evolutionary history and adaptive mechanisms. Over the past few decades, researchers have made significant progress in understanding the genomic structures of microsporidia. Through genome sequencing, scientists have gained insights into the biology, evolution, and pathogenicity of microsporidia [Keeling et al., 2005, Weiss and Reinke, 2022].

1.1.4 Microsporidia in invertebrates

Microsporidia can infect a wide range of invertebrates, including nematodes [Troemel et al., 2008], amphipods [Bojko and Ovcharenko, 2019], planktonic crustaceans such as *Daphnia* [Refardt et al., 2002] and mosquitoes [Becnel et al., 2005]. Microsporidia are transmitted through multiple routes, depending on the host species and environmental status. Transmission can occur orally by ingesting infected

spores, direct contact with infected individuals or contaminated environments, or from an infected mother to their offspring [Fast and Keeling, 2005, Wittner and Weiss, 1999]. In invertebrates, microsporidia can be transmitted vertically from mother to offspring through the transovarial route, leading to a high and sustainable infection prevalence within host populations [Becnel and Andreadis, 1999]. Horizontal transmission, where the parasite spreads from one individual to another, typically through environmental spores, is also standard, particularly in species infecting arthropods and fish [Fast and Keeling, 2005, Pan et al., 2018].

1.1.4.1 Development of new and innovative malaria control tools

Malaria cases have been on a steady increase since 2016, with a notable 247 million cases reported in 2021 alone. A total of 619,000 malaria-related deaths were recorded, 96% of which were from sub-Saharan Africa [World Health Organization, 2022]. The major impediments to global malaria control have included the development of insecticide resistance in the *Anopheles* vector, *Plasmodium* drug resistance, poor detection of the minor malaria parasites, and the recent invasion of a new vector, *An. stephensi*, in the Horn of Africa and urban environments [Ojuka et al., 2015, Evaluation, 2016, World Health Organization, 2022, Whittaker et al., 2023, Ochomo et al., 2023]. Controlling vector populations has been the most efficient way of controlling malaria transmission [Hemingway, 2014]. However, as efforts to control malaria have plateaued, there is a growing concern about the need to innovate vector control methods to overcome resistance and improve the efficiency of the current tools. Much of this uses biological control tools to complement the current vector control strategies. This can be a promising way to address the challenges of malaria control and potentially improve control efforts.

1.1.4.2 The mosquito microbiome and impact on vectorial capacity

Significant reductions in *Plasmodium* numbers within the mosquito during the parasite's life cycle potentially affect malaria transmission [Smith et al., 2014]. Interventions targeting these bottlenecks, such as transmission-blocking vaccines, genetic modifications of vectors, and integrated vector management strategies, could significantly reduce the spread of malaria. The tripartite interaction be-

tween *Plasmodium*, mosquitoes, and the host microbiome is an essential factor in malaria transmission efficiency. The mosquito natural microbiome has been reported to have a crucial impact on its vectorial capacity and pathogen transmission [Dennison et al., 2014, Hegde et al., 2015, Cansado-Utrilla et al., 2021]. Microbes such as bacteria, viruses, fungi, and microsporidia can modulate vector competence [Dada et al., 2021] influencing vector density, biting rate, vector competence, and pathogen extrinsic incubation period [Boissière et al., 2012, Dong et al., 2009, Engel and Moran, 2013, Minard et al., 2013, Gendrin et al., 2015, Gao et al., 2020, Djihinto et al., 2022]. Symbionts like bacteria and fungi have been highlighted as potential tools for controlling mosquito-borne diseases. Bacteria species such as Wolbachia [Aliota et al., 2016, Ant et al., 2018, Slatko et al., 2014], Spiroplasma [Herren and Lemaitre, 2011], Serratia [Bando et al., 2013, Jupatanakul et al., 2020], Rickettsia [Lukasik et al., 2013] and Elizabethkingia [Garay et al., 2016] have been shown to contain protective capacities in their respective insect hosts. Wolbachia has been extensively studied and applied in the control of arbovirus transmission in *Aedes* [Ant et al., 2018], highlighting the potential of tapping into the natural microbiome of mosquitoes to control vector-borne disease transmission.

Fungi have been shown to dictate insect host behavior [Shang et al., 2015]. Entomopathogenic fungi have also been identified as promising malaria biocontrol agents due to their ability to infect and kill mosquitoes at various life stages [Scholte et al., 2004, Nattoh et al., 2021]. Reducing mosquito populations can, however, have crucial ecological risks. Reducing mosquito populations could disrupt food chains, leading to unforeseen consequences for predator species that rely on mosquitoes as food sources. An ecosystem approach incorporating ecological thinking into the development and implementation of vector control activities is essential to mitigate these risks and ensure sustainable and effective management solutions, such as using symbionts with minimal effects on mosquito populations.

1.1.4.3 Microsporidia in vector control

Microsporidia have garnered significant attention from vector biologists due to their potential as biological control agents for managing vector populations and

disease transmission. Microsporidia have garnered significant attention from vector biologists due to their potential as biological control agents for managing vector populations and disease transmission [Andreadis, 1990, Bjørnson and Oi, 2014]. Arthropod pathogens are potentially crucial biological control agents in pest management and are essential factors that regulate the natural populations of their hosts [Andreadis, 1990]. Their ability to selectively target and infect specific vectors, such as mosquitoes or ticks, makes them an attractive option for controlling the transmission of vector-borne diseases such as malaria, dengue fever, or Lyme disease. Therefore, the search for an alternative vector control method, including biological control, has attracted the attention of many recent studies [Caragata and Walker, 2012, McGraw and O'Neill, 2013]. Microsporidia have shown promise as biocontrol agents against mosquito vectors [Murareanu et al., 2021]. Furthermore, microsporidia infections in mosquitoes have been found to alter their blood-feeding behavior, kill larvae and pupae, reduce adult longevity and fecundity, and hinder the development of *Plasmodium* parasites [Andreadis, 2007, Becnel et al., 2005, Herren et al., 2020]. A wide range of microsporidia have been found to infect mosquitoes [Trzebny et al., 2020], opening up the possibility of harnessing these organisms as biocontrol agents against these vectors to prevent the spreading of diseases like malaria.

1.1.4.3.1 Microsporidia of mosquitoes and their role in the control of vector-borne diseases

Microsporidia infect a wide range of mosquito species and have diverse effects on their hosts depending on their species (Table 1.1). Mosquito-microsporidia interactions are multifaceted, involving host immune responses crucial for controlling infections. These defense mechanisms against microsporidia include ubiquitin-mediated resistance and the production of reactive oxygen species in the hosts [Pan et al., 2018, Szumowski and Troemel, 2015]. Understanding these interactions is essential for developing strategies to manage mosquito populations and control the spread of mosquito-borne diseases. Research on mosquito immune response to microsporidia is still developing, with the application of genomic approaches taking the front line in understanding the genetic and molecular basis of

Table 1.1: Overview of Mosquito-infecting Microsporidia as highlighted in [Andreadis, 2007].

Microsporidia species	Mosquito species	Effect on Host	Transmission Mode
<i>Edhazardia aedis</i>	<i>Aedes</i>	Pathogenic	Horizontal & vertical
<i>Vavraia culicis</i>	<i>Culex</i>	Pathogenic	Horizontal & vertical
<i>Amblyospora spp.</i>	<i>Aedes</i> and <i>Culex</i>	Pathogenic	Horizontal & vertical
<i>Anncaliia algerae</i>	<i>Aedes</i> and <i>Culex</i>	Pathogenic	Horizontal & vertical
<i>Hazardia muelleri</i>	<i>Anopheles</i> and <i>Culex</i>	Pathogenic	Horizontal
<i>Culicospora spp.</i>	<i>Anopheles</i> and <i>Culex</i>	Pathogenic	Horizontal & vertical
<i>Parathelohania spp.</i>	<i>Anopheles</i> and <i>Aedes</i>	Pathogenic	Horizontal & vertical

susceptibility and refractoriness in mosquitoes against other pathogens aside from microsporidia [Zheng, 1997].

Multiple methods are used in the taxonomy of microsporidia. Both morphological and molecular approaches are used to classify mosquito-infecting microsporidia. Moreover, these obligate intracellular parasites are classified based on their life cycles and host-parasite relationships, with some species exhibiting simple life cycles and broad host ranges such as *Anncaliia algerae* and *Vavraia culicis*, while other mosquito microsporidia such as *Edhazardia aedis* have complex life cycles with high host specificity [Becnel et al., 2005]. The gold standard molecular method of classifying these microsporidia using the characterization of ribosomal RNA (rRNA) is currently being complemented with structural and multiple molecular markers to improve the classification of the lineage and functional diversity of microsporidia [Vossbrinck et al., 2004, Bojko et al., 2022]. Most importantly, the availability of multiple microsporidia genomic data has enabled the use of overall genome relatedness indices (OGRI) such as average nucleotide identity (ANI) to improve taxonomic evaluation of these taxonomic nomads [Haag and De Albuquerque, 2022].

1.1.5 *Microsporidia* sp. MB as a potential vector control tool

Microsporidia sp. MB is a unique species found in multiple *Anopheles* hosts across Africa, including *An. gambiae sensu lato*, *An. funestus* and *An. coluzzii* [Herren et al., 2020, Nattoh et al., 2023, Akorli et al., 2021, Moustapha et al., 2024]. It has been established that this organism has a strong malaria transmission-blocking phenotype in *An. arabiensis* characterized by its refractory effect on oocyst development in the mosquito [Herren et al., 2020]. The symbiont is both vertically (transovarial) [Herren et al., 2020, Makhulu et al., 2023] and sexually transmitted from male to female [Nattoh et al., 2021, 2023], ensuring its maintenance in host populations and thus a potentially effective and sustainable biocontrol tool [Bukhari et al., 2022]. The vertical transmission of *Microsporidia* sp. MB in *Anopheles arabiensis* is an important feature in its spread across the mosquito

population by influencing generation-to-generation transmission, enhancing its establishment, population-level spread, and maintenance within host populations. This species is also avirulent in mosquitoes. It exhibits little to no effect on host fecundity and lifespan, a crucial feature for its long-term establishment and maintenance within mosquito populations [Herren et al., 2020, Nattoh et al., 2023]. *Microsporidia* sp. MB shows specific tissue tropism and localization patterns in *An. arabiensis*, particularly in male and female reproductive tissues, linking this to their sexual and maternal transmission routes [Makhulu et al., 2023]. The complete infection dynamics of this microorganism are yet to be fully understood. Studies on the impact of blood-feeding on *Microsporidia* sp. MB infection levels describe a notable increase in the intensity of the symbiont infection in female ovaries, suggesting a possible correlation between symbiont levels, oogenesis, and blood-feeding [Makhulu et al., 2023]. Mathematical modeling analysis has also been conducted to study the dynamics of *Microsporidia* sp. MB-infected mosquito spread in wild populations, considering factors such as transmission efficiencies, mating behaviors, mosquito development stages [Charlene N et al., 2023, Mfangnia et al., 2023]. The study highlights the potential coexistence of infected mosquitoes in the wild and predicts the expected outcomes of introducing these mosquitoes in the field as a malaria control strategy. Despite the extensive research conducted on *Microsporidia* sp. MB, a formal species name has not yet been assigned to the organism. Although *Microsporidia* sp. MB is now one of the most extensively characterized microsporidian species, formal classification is a rigorous process requiring consensus among taxonomists and other experts in the field. Ongoing efforts with collaborators aim to fully classify this microsporidia and assign a proper species name, ensuring that the taxonomic nomenclature aligns with its genomic and biological characteristics. Overall, *Microsporidia* sp. MB shows unique attributes that make it an effective and sustainable biocontrol tool. Extensive research on its transmission and infection dynamics has been conducted; however, much has yet to be researched on the biology of these intricate microsporidia.

1.2 Knowledge gap

Understanding the genome of microsporidia is important for discerning the genetic basis of different species and their unique phenotypic effects on their hosts. This dependence on the host makes microsporidia an interesting model for studying the intricate interactions between parasites and their hosts. In addition, the high levels of intraspecific genomic diversity observed in microsporidia highlight the potential for studying the factors driving this variation and how it contributes to the evolution and adaptation of these organisms [Huang et al., 2023]. *Microsporidia* sp. MB is a fairly new species isolated from *Anopheles arabiensis* and shown to have an effective capacity to block malaria development [Herren et al., 2020]. One of the primary challenges impeding the study of this species is the lack of genome sequence data for this microsporidia. This makes it difficult to understand the genetic basis of their unique biology and symbiosis with the mosquito host. The variations in microsporidia genomes, such as their extreme reduction in size and gene numbers, indicate a heavy reliance on their hosts for biochemical processes [Peyretailade et al., 2011]. Understanding these interactions is essential for developing effective *Microsporidia* sp. MB molecular experiments. Information about the specific genes and pathways involved in the host-parasite interaction is currently unavailable. *Microsporidia* sp. MB has been found in different species of *Anopheles* across Africa [Herren et al., 2020, Akorli et al., 2021, Nattoh et al., 2023, Moustapha et al., 2024]. However, the possibility of strain variations across different regions and host species has not been investigated. The availability of a reference genome for this microorganism will provide the needed resources to conduct strain variation analysis and determine the probable impact of different strains on the efficiency of the development of blocking *Plasmodium*. Lastly, there is a need to explore the potential of microsporidia as biocontrol agents against other pathogens, which requires a better understanding of their ecology and interactions with other organisms. This will be possible with information on their genetic makeup.

1.3 Research aim and objectives

1.3.1 Research aim

Microsporidia are unique organisms that elicit diverse effects on their hosts. The discovery of *Microsporidia* sp. MB and its promising future as a sustainable bio-control agent have raised global interest among malaria scientists. However, little is known as to how this symbiont blocks malaria development within the mosquito and its intricate relationship with its host at a molecular level. Presently, the knowledge of *Microsporidia* sp. MB has been limited to experimental analyses of the prevalence, transmission routes, infection dynamics, and parasite-blocking characteristics of this symbiont. The absence of information on its genetic makeup has impeded more extensive studies of *Microsporidia* sp. MB on a molecular level. This thesis aims to obtain insights into the basic biology of the malaria-transmission-blocking symbiont of *Anopheles arabiensis*, *Microsporidia* sp. MB by sequencing and analyzing the structure and content of its genome using bioinformatics tools, providing a reference that would act as a backdrop to future molecular experiments to understand the organism fully.

1.3.2 Main objectives

1. To develop a computational annotation pipeline to describe the structure and function of hypothetical microsporidia proteins using *Vittaforma corneae* ATCC 50505 as a model by conducting sequence homology, physicochemical, phylogeny, domain, and motif analyses.
2. To sequence and assemble a reference genome of *Microsporidia* sp. MB and describe its general genomic features.
3. To study the function and structure of the *Microsporidia* sp. MB proteome using computational tools described in objective one above and characterize various biological pathways in this species.

To this end, this study successfully describes the assembly, structural, and functional annotation of the first reference genome of *Microsporidia* sp. MB and

highlights the unique biological processes occurring in the symbiont.

Chapter 2

Structural and Functional Annotation of Hypothetical Proteins from *Vittaforma* *corneae* ATCC 50505 Using in silico Approaches

2.1 Chapter Overview

Microsporidia are organisms with compact genomes due to the selective loss of genes during evolution, rendering them entirely dependent on the hosts for survival. Characterizing microsporidia protein function and structure has provided valuable insights into the mechanisms of infection and pathogenesis caused by *Vittaforma corneae* and other related microsporidian species. These studies have shed light on the specific proteins involved in host cell invasion, replication, and immune evasion, as well as the structural characteristics that contribute to the unique biology and survival strategies of these parasites. Despite the small number of genes, a high percentage of the genes in microsporidia genomes code for proteins whose functions are unknown. These unknown proteins are referred to as hypothetical proteins (HPs). Computational annotation of HPs has become a more efficient and cost-effective alternative to experimental investigation. This

chapter highlights the development of a robust bioinformatics annotation pipeline of HPs from *V. corneae*, a microsporidian that causes ocular infections in immunocompromised individuals. The process involved retrieving sequences and homologs, performing physicochemical characterization, protein family classification, motifs and domains analysis, protein-protein interaction (PPI) network analysis, and homology modeling using various online resources. Consistent findings across platforms demonstrate the accuracy of annotation using these *in silico* methods. Of the 2,034 HPs retrieved from online databases, 162 were fully annotated in this study, with most categorized as binding proteins, enzymes, or regulatory proteins. The functions of several HPs from *V. corneae* were accurately inferred, improving our understanding of microsporidian HPs despite challenges related to the obligate nature of microsporidia, the absence of fully characterized genes and the lack of homologous genes in other systems.

2.2 Keywords

Hypothetical Proteins; Microsporidia; Computational Annotation; Functional and Structural Annotation; Protein Function Prediction; Physicochemical Properties; Homology Modeling; Motif Analysis; Conserved Domains.

2.3 Introduction

Vittaforma corneae ATCC 50505, also known as *Nosema corneum*, is a microsporidia that infects the corneae of mammals, causing keratoconjunctivitis [Moshirfar et al., 2020]. It is a vital microsporidian public health concern due to its potential to cause severe eye infections and visual impairments, particularly in individuals with compromised immune systems [Deplazes et al., 1998]. The organism is found to be ubiquitous in the environment, and various factors can predispose individuals to *V. corneae* infection, including exposure to soil or mud in water, dust particles, use of contact lenses, insect bites, bathing in unclean river waters, and even exposure to hot springs [Chen et al., 2018, Leroy et al., 2018].

2.3.1 *V. corneae*-host cell interactions

V. corneae induces the formation of multi-nucleate structures in host cells, subverting the host cell cytoskeleton and volume control [Mittleider et al., 2002, Franzen and Salzberger, 2008]. This interaction is particularly relevant in the cornea, where the pathogen can cause infection when the ocular surface is perturbed [Moshirfar et al., 2020]. The cornea’s defense mechanisms, including cell-cell interactions, are crucial in preventing and treating such infections. *V. corneae*-host cell interactions play a crucial role in the pathogenesis of infections caused by this microsporidian parasite. These interactions involve the attachment, invasion, and manipulation of host cells by *V. corneae*, leading to various pathological manifestations. Understanding the mechanisms of *V. corneae*-host cell interactions are essential for developing effective therapeutic interventions and preventive strategies by uncovering key molecular mechanisms involved in the infection process, replication, and immune evasion of this microsporidia.

2.3.2 Hypothetical proteins (HPs)

HPs present a unique challenge since no experimental data is available on their *in vivo* expression. Studying the structure and function of these proteins can provide valuable information about their roles in metabolic pathways and help decipher their significance in the biology and infection phenotype of microsporidia [Eisenstein et al., 2000, Galperin, 2001, Sivashankari and Shanmughavel, 2006, Desler et al., 2009, Shahbaaz et al., 2013]. Numerous studies on the use of fast, cost-effective, and accurate computational annotation methods have been described for proteins from human pathogens [Mazandu and Mulder, 2012, Shahbaaz et al., 2013, Kumar et al., 2014] showing the capacity of these tools in providing important and accurate information on unknown proteins from these microorganisms. Key applications of computational annotations of HPs also include a recent report on novel abiotic stress proteins in *Triticum aestivum* [Gupta et al., 2018] and the identification of new drug targets [Omeershffudin and Kumar, 2019].

2.3.3 Distribution of HPs in microsporidia proteomes

As minimalist obligate intracellular eukaryotes, microsporidia have lost a lot of genes over time. The microsporidia proteome consists of over 50% unassigned proteins, which are believed to play an essential role in these microorganisms' physiological and biochemical processes [Texier et al., 2005]. Annotating these proteins is crucial in exploring their basic functions and interactions with other proteins. *V. corneae* has a relatively small genome of 3.2 Mb [Mittleider et al., 2002] spanning 2,237 proteins. A total of 2,034 of these proteins are labeled as HPs [Mittleider et al., 2002]. In the 2002 study of *V. corneae*, 90% of the proteins were classified as hypothetical due to the limited annotation resources available at the time. This was a common occurrence in early genome sequencing projects, particularly for organisms like microsporidia, where much of the protein-coding potential remained poorly understood. Since then, genome annotation has progressed significantly, driven by efforts such as the Broad Institute's genome assembly. Many of these previously termed "hypothetical" proteins are now referred to as "uncharacterized" proteins, reflecting ongoing research aimed at uncovering their functions. While there has been a shift from incomplete annotation toward a more refined understanding, many proteins in *V. corneae* still require full characterization.

2.3.4 Computational approaches in protein functional and structural annotation

Recently, the availability of large amounts of genomic data has made it imperative to employ bioinformatics tools for sequence analysis. This helps identify protein structure and function [Zhang and Wang, 2023]. Different bioinformatics tools and resources have been developed to annotate protein structure and function, including online protein databases, domain and motif analysis, and 3D protein modeling. A recent study by [Mascarenhas dos Santos et al., 2022] provides a strong example of using structural homology-based approaches to annotate microsporidia proteins, revealing the presence of key DNA repair machinery pre-

vously thought to be absent from these organisms.

2.3.4.1 Protein databases

Protein databases are essential for understanding protein functions and their roles in biological systems using bioinformatics methods and computational resources. These repositories contain information on protein structure, domains, three-dimensional conformations, molecular functions, post-translational modifications, and interactions used to explore the biological characteristics of proteins [Mehmood et al., 2014, Ramazi and Zahiri, 2021]. These online resources can identify relationships within a genome or across different species by comparing proteins and protein families [Mehmood et al., 2014]. Protein structure and function can also be predicted using secondary databases, which annotate and provide predictions based on experimental data. These protein databases are constantly being curated, as they contain not only a collection of data but also bioinformatics infrastructure.

2.3.4.2 Protein homology search

Protein homology search is a bioinformatics approach used to identify evolutionary-related proteins based on sequence similarity, providing information on protein structure and function [Margelevičius and Venclovas, 2010]. Sequence-based homology search methods such as BLAST are the benchmark in protein homology studies and have significantly improved over time. However, remote homologies often go undetected when solely relying on sequence homology, compared to using protein spatial structures [Suplatov et al., 2016]. Therefore, methods incorporating non-sequence-based information can complement the sequence-based method to provide insights into homology details that the latter miss [Memišević et al., 2010]. DNA-protein homology search methods have also been developed to detect protein fossils with improved sensitivity and speed compared to the traditional BLAST methods. Improvements in the precision of sequence homology prediction have also included incorporating biomedical literature into the PSI-BLAST algo-

rithm [Chang et al., 2000]. HMMER software has also improved these homology searches using stringent profile hidden Markov models (HMMs) [Finn et al., 2011]. Prediction of protein families from sequences with no known homologs is also possible using support vector machines (SVM) [Han et al., 2004]. Protein language model search (PLMSearch) has also significantly improved sensitivity over other sequence search methods by using deep representations from pre-trained protein language models, identifying homology even with low sequence similarities [Liu et al., 2023]. Integrating protein sequence-based methods with additional data sources such as structural information, PPI network analysis, advanced computational methods, and literature to improve detection and annotation of homologous sequences [Keskin et al., 2016].

2.3.4.3 Multiple sequence alignment (MSA) and phylogenetic analysis

MSA is an essential process in bioinformatics that aims to align three or more biological sequences to identify conserved regions that might provide insights into the structural, functional, and evolutionary relationships of these proteins [Ozer et al., 2010, Reddy and Fields, 2022]. Accurate phylogenetic analysis relies heavily on the quality of MSA, with errors in MSA leading to errors in the inferred phylogenetic tree [Yue et al., 2009]. Although traditional MSA algorithms are accurate, they are usually computationally exhaustive. This has led to the development of various heuristic and meta-heuristic approaches that can achieve near-optimal alignments more efficiently [Issa and Hassanien, 2020, Reddy and Fields, 2022]. MSA techniques have drastically improved recently with the development of algorithms that integrate both the alignment process and phylogenetic tree construction to simultaneously optimize both [Yue et al., 2009]. For example, the algorithm described by Yue [Yue et al., 2009] aligns sequences while searching for the phylogenetic tree that minimizes the alignment score, showing promising results in both alignment quality and phylogeny accuracy. In situations where sequence identity is low, structural data and network properties have been utilized to enhance MSA methods by identifying domains and similar structures [Ozer et al., 2010]. Additionally, protein profiles and domain information, such as PFam

annotations, have improved the precision and robustness of DNA multiple alignments [Balech et al., 2018]. Bioinformatics tools have been designed to facilitate MSA analysis, such as the "msa" package in R [Bodenhofer et al., 2015], Geneious (a simple graphical user interface for alignment and visualizations of sequences) [Kearse et al., 2012] and Molecular Evolutionary Genetics Analysis (MEGA X) [Kumar et al., 2018], among others. Computational tools for MSA and phylogeny studies, like many bioinformatics tools, are continuously being updated to ensure high alignment quality and computational efficiency while being user-friendly.

2.3.4.4 Protein physicochemical analysis

Protein physicochemical analysis includes characterization of protein hydrophobicity (affinity to water), polarity, and charge, all fundamental properties in understanding the molecular and biochemical mechanisms of proteins in-depth [Cortés and Aguilar-Ruiz, 2011]. Computational physicochemical properties of proteins provide insights into the structure, function, and interactions of proteins. The Expert Protein Analysis System (ExPASy) ProtParam tool is a bioinformatics resource that facilitates the computation of various physical and chemical properties associated with protein sequences [Gasteiger et al., 2005a]. These parameters include the molecular weight, theoretical isoelectric point (pI), amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index, and grand average of hydropathicity (GRAVY). This tool analyzes primary protein sequences as input and can provide information about the protein structure and function. These parameters are important in understanding protein behavior in biological systems, guiding experimental designs of these proteins. The instability index predicts the stability of a protein in a test tube, where a value of more than 40 indicates that the protein may be unstable. The molecular weight is the sum of the atomic weights of all atoms in a molecule expressed in Daltons and is used to evaluate protein purity and concentration. The pI is the pH at which a protein carries no net electrical charge. This parameter is essential in protein separation and purification experiments such as isoelectric focusing. The amino acid composition measures the percentage of each amino

acid in the protein sequence and can be used to assess the protein structure and function. On the other hand, atomic composition is the proportion of each atom present in the protein and helps identify the elemental analysis of the protein. The extinction coefficient measures the amount of light absorbed by a protein at a specific wavelength, which is important in determining protein purity and quantity. The estimated half-life is the time taken for half of the protein molecules to degrade. This is usually expressed in hours to evaluate protein turnover. The aliphatic index determines protein thermostability by measuring the relative volume of aliphatic side chains. These include glycine, valine, alanine, isoleucine, and leucine. Finally, GRAVY measures the hydrophobicity or hydrophilicity of a protein and is indicative of membrane protein topology, biochemical properties, and possible PPIs [Gasteiger et al., 2005a].

2.3.4.5 Protein motif discovery

Protein motif analysis focuses on identifying recurring patterns within protein sequences, which could indicate structural and functional properties [Glasgow et al., 1999, Sahu et al., 2012]. It combines computational methods with biological knowledge to identify and explore complex patterns underlying protein structure and function. The constant improvement of these algorithms and tools has enhanced the ability to accurately identify and analyze these motifs, contributing to a deeper understanding of protein biology [Glasgow et al., 1999, Sahu et al., 2012]. These structure-based or sequence-based motifs can be pivotal in understanding sequence function and evolution and applied in drug design studies [Elayaraja et al., 2012]. Systemic analysis and classification of proteins based on shared motifs is essential in investigating extensive sequence data generated by genome sequencing. Several motif discovery databases and algorithms have been designed and improved over time, with some tools specifically developed to predict protein structural features and construct 3D models from protein sequences [Mooney et al., 2011]. These computational tools and algorithms are designed to identify both known and unknown motifs, with some identifying these patterns in unaligned sequences [Bailey, 2008, Mooney et al., 2011, Semwal et al., 2022]. K-mer-based methods, including

Multiple Expectation-maximization for Motif Elicitation (MEME), GLAM2, and Pr[m], are used to detect statistically significant ungapped motif patterns (short contiguous amino acid sequences of size "k") [Bailey et al., 2006, Semwal et al., 2022]. Cluster algorithms such as K-means and rough K-means are useful in finding conserved protein patterns that are not easily detectable during MSAs in a set of sequences. These algorithms cluster similar sequences based on distance or similarity measures. The clusters formed can provide insights into the recurring patterns or motifs in a set of protein sequences. In contrast, the structural similarity of the discovered clusters can provide information about the conservation of these motifs across different organisms or protein families [Elayaraja et al., 2012]. The *ab initio* motif finder NestedMICA can sufficiently identify short protein signals even when these motifs are only present in small sequences without prior knowledge of the motif structure or sequence. NestedMICA uses a nested iterative strategy to identify motifs at different levels of conservation and can identify these patterns in noisy or incomplete datasets [Doğruel et al., 2008].

2.3.4.6 Conserved domains analysis

Proteins conserved domains refer to the recurring segments or patterns found within protein sequences that can independently fold, function, and evolve. These domains maintain a specific structure and function across different proteins and species. They play a crucial role in protein function and can provide important insights into the evolutionary relationships between different organisms [Wang et al., 2021]. The Conserved Domain Database (CDD) is a primary resource that provides annotations of protein sequences with conserved domain footprints, which are crucial for elucidating protein function, classification, and evolutionary mechanisms [Yang et al., 2020]. Conserved domain profiles, when combined with sequence profiles, can enhance the accuracy of protein secondary structure prediction, as demonstrated by an increase in prediction accuracy on the RS126 data set [Woo et al., 2005]. Besides identifying basic protein parameters, the analysis of conserved domains also extends to identifying conserved sites, predicting folding, structure, and function, and constructing phylogenies [Zhang and Wang, 2023].

Protein functional sites (PFSs) and protein-ligand interactions (PLIs) can also be identified using domain analysis. These protein features are key to understanding biochemical reactions within the organism. A function- and interaction-annotated protein domain profile database (fiDPD) has been used to predict PFSs and PLIs accurately [Ming et al., 2017]. Conserved domain analysis forms a basis for further discovery and experimental validation in protein research and can be particularly useful in predicting the functions of proteins whose exact roles are yet to be determined [Wang et al., 2021, Yang et al., 2020].

2.3.4.7 Homology modeling and three-dimensional (3D) structural analysis

Protein homology modeling is a computational technique used to predict the 3D structure of a protein based on its sequence similarity to proteins with known structures [Xiang, 2006, Gupta et al., 2016]. The homology modeling process involves aligning protein sequences with template structures, followed by fold assignment, model building, and qualitative evaluation to generate high-quality 3D models of target proteins [Marti-Renom et al., 2002, 2003]. This technique has been successfully applied in drug design studies and virtual screening applications by providing insight into the spatial arrangement of residues [Gupta et al., 2016, Kopp and Schwede, 2004]. Homology modeling has become an accurate computational technique for creating structural models that can be as accurate as low-resolution experimental data, especially when the sequence identity to the template is over 30% [Xiang, 2006]. Comparative homology modeling has become a fundamental aspect of protein function annotation and structural genomics, primarily due to the widespread availability of automated pipelines and model repositories [Hatherley et al., 2016, Ronneberger et al., 2021, Kopp and Schwede, 2004]. Homology modeling has significantly helped fill the gap between the abundance of protein sequences and the scarcity of their corresponding structures despite the lack of structural information for many protein sequences, making this contribution valuable to molecular biology.

2.4 Methodology

Functional and structural annotation of HPs was done in two primary phases (Figure 2.1). Default parameters were used for each tool unless otherwise noted.

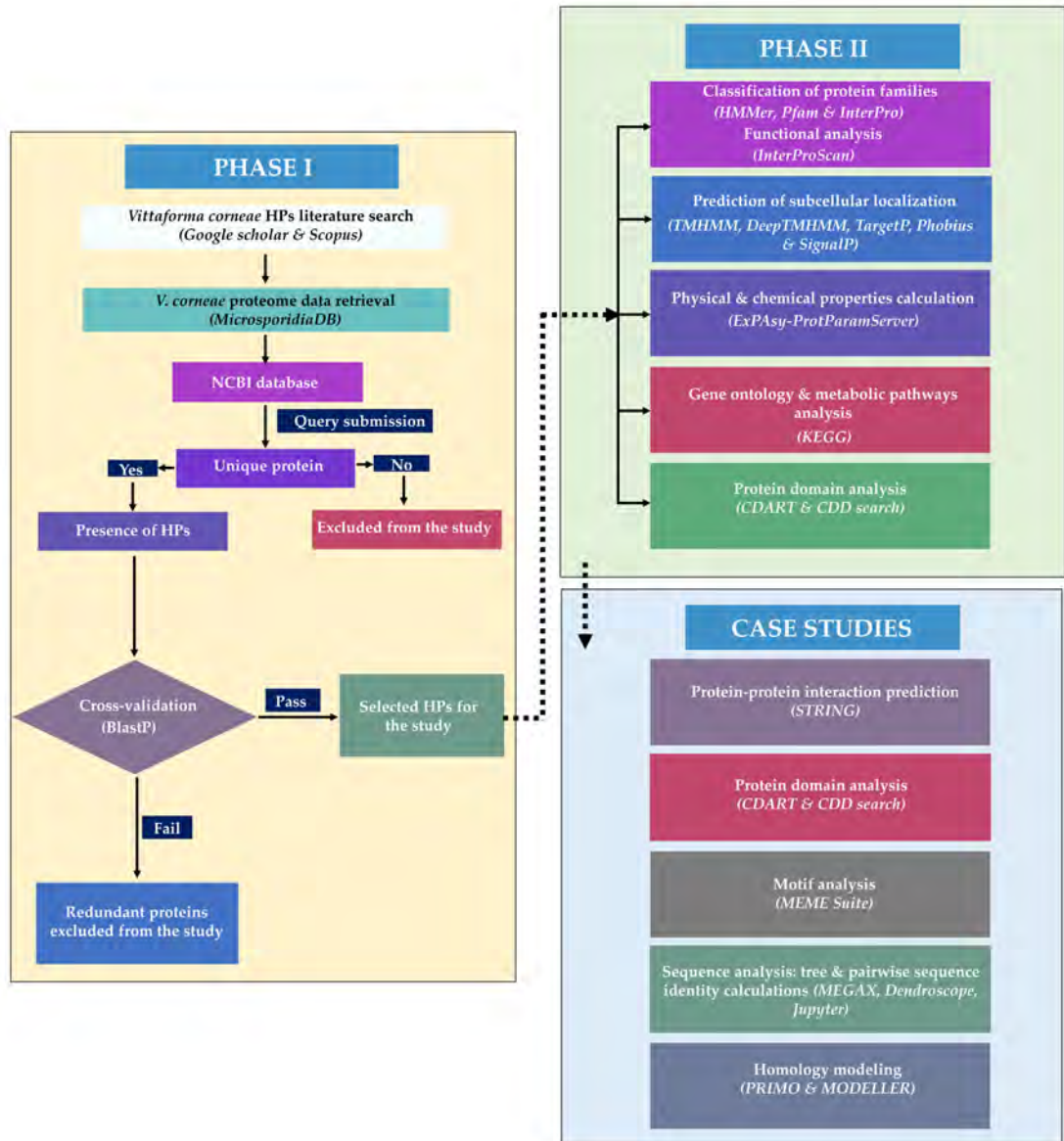


Figure 2.1: A summary of the computational workflow designed in this study to annotate *V. corneae* HPs. The pipeline highlights two major phases: PHASE I. Sequence retrieval, similarity search, and test for homology. PHASE II. Diverse functional annotation using different tools to predict protein families, localization, motifs and domains identification, chemical properties of the HPs, sequence identity calculations, homology modeling, and phylogenetic analysis.

2.4.1 Phase I

2.4.1.1 Literature search of *V. corneae* proteins

Using the keywords " *Vittaforma corneae* hypothetical proteins," " *Vittaforma corneae* unknown proteins characterization," and "Function and structure prediction of *Vittaforma corneae* unknown proteins," an extensive literature search on *V. corneae* proteins was conducted on Scopus [Burnham, 2006] and Google Scholar [Jacsó, 2005] databases [accessed on December, 2022]. Following this search, no relevant research specific to *V. corneae* was identified. Following this literature review, this study was conceptualized, and the annotation framework was designed to characterize the HPs from this microsporidia (Figure 2.1 above). The designed pipeline was divided into two primary phases: PHASE I encompassed identifying and retrieving unique HPs from online databases and a homology search on NCBI; PHASE II detailed the key tools used for the functional and structural annotation of the selected proteins. These annotations included protein family classification, characterization of the HP physicochemical properties, prediction of membrane-spanning proteins and presence of signal peptides, identification of PPI networks, and exploring the associated metabolic pathways. A summary of the tools and online resources used in this study is highlighted in the List of Tools.

2.4.1.2 Sequence retrieval

The proteomic data of *V. corneae* was retrieved from the Microsporidia Database microsporidia_odb10 v10 (<https://microsporidiadb.org/micro/app/search?q=+Vittaforma+corneae+ATCC+50505>) in FASTA format [Aurrecoechea et al., 2011] [accessed on December 8, 2022]. The proteome contained 2237 proteins, 90.97% ($n = 2034$) labeled as HPs. To retain only the HPs from the FASTA file, a customized Python script was developed using Python v3.8 [Van Rossum and Drake, 2000]. The script worked by checking all the sequence header descriptions for the presence of the term "HYPOTHETICAL" and subsequently discarding all sequences that did not meet this criteria. The sequence further ensured that all the retained HP sequences were unique by cross-referencing each sequence against each other, retaining only

the non-redundant ones. Following this rigorous filtering process, a refined subset comprising 164 HPs was ultimately used in this study. This thorough selection and filtering procedure was followed to improve the precision and relevance of the data being examined.

2.4.1.3 Homology search

Homology search of the initial 2034 HPs was done using the remote BLAST+ command-line tools against the NCBI BLASTp standard non-redundant protein sequences (nr) database [Johnson et al., 2008, Sayers et al., 2021] [accessed in December 2022]. This step aimed to confirm that these proteins possessed known functions. The search was carried out with default parameters (E-value = 0.0001, bit-score greater or equal to 200), limiting the search to a maximum of 10 target sequence hits. The BLAST+ command-line code options included: *"blastp -db nr -remote -prot"*. Following this initial search, 627 of the HPs showed no similarity to any sequences in the database, while 677 had relevant hits. To extend the depth of the search for distantly related sequences, the Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) was used [Kumar et al., 2012] [accessed between 8-16 December 2022]. PSI-BLAST uses an iterative approach that uses position-specific matrices (PSSMs) to score matches between the query and the database, comparatively similar to the algorithm used by the HMMER (jackhammer) program [Eddy, 2020]. Subsequent analyses excluded 627 sequences that did not have detectable homologs in either BLASTp or PSI-BLAST. These sequences were excluded because they either lacked homologous counterparts altogether, were only homologous to HPs from other microsporidian species or exhibited insufficient sequence similarity and coverage to any known sequences in the database. Filtering these HPs based on these criteria was important since the functional and structural annotation was based on inference from the information available on well-characterized microsporidian proteins.

Reverse BLAST analysis was further done to confirm true homology (filter out sequences that may have resulted by chance) and refine the dataset. An in-house Python script, developed using Jupyter Notebook v1.2.6 [Van Rossum and Drake,

2000, Silaparasetty, 2020], executed a remote command-line BLAST+ on the HP homologs identified in the top 10 hits described above and filtered the output hits. The stringent criterion for classifying HPs as true homologs included confirming that the HP appears among the top 3 hits with relatively good percent identity (greater than 40%), high query coverage (greater than 50%), and a low E-value (less than 0.005). The resulting sequences identified as true homologs to their characterized counterparts ($n = 164$) were subsequently used downstream in silicon characterization steps. Figure 2.2 below shows a schematic representation of the reverse BLAST technique.

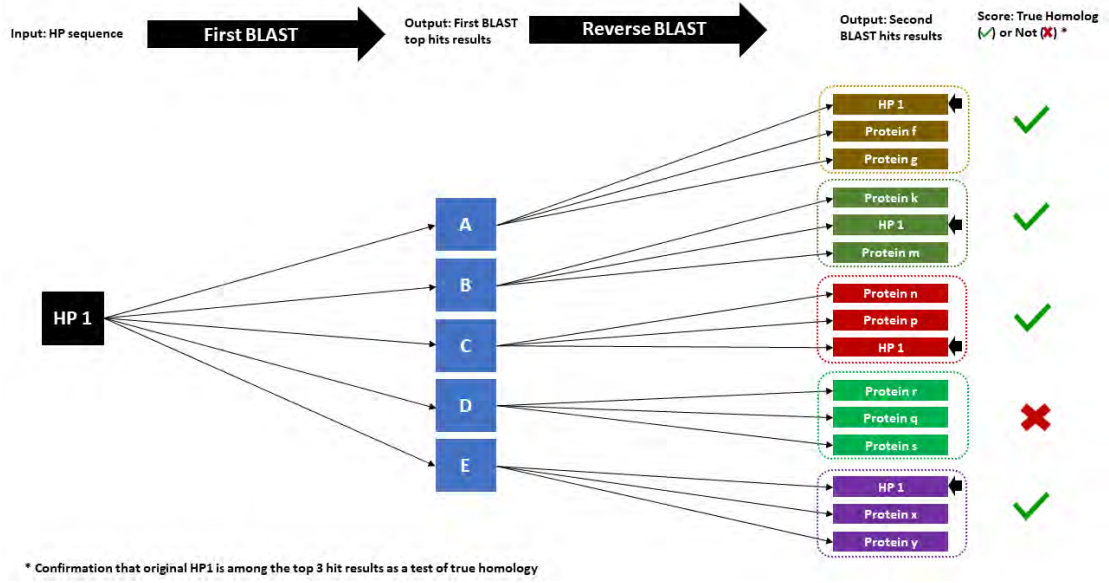


Figure 2.2: A schematic representation of the reverse-blast method of testing true homology. To confirm the true homology of the HPs, the top 10 BLASTp hits were selected for each protein. The identification of true homologs was based on the appearance of the HP in the top three hits with high coverage and percent identity from the sequences analyzed. .

2.4.2 Phase II

2.4.2.1 Protein family classification

PFam v34 [Mistry et al., 2021] and InterPro [Blum et al., 2021] embedded in the HMMer web server [Potter et al., 2018] were used to identify protein family categories of HPs using default parameters. The top hit with statistically significant search output was used to categorize the protein families (significance E-values for hits = 0.03 and sequence = 0.01). PANTHER [Mi et al., 2021], PRINTS [Attwood, 2002], PIRSF [Wu et al., 2004], and GENE3D [Lees et al., 2014] were also used to complement the Pfam and InterPro annotations.

2.4.2.2 Sub-cellular localization and signal peptide prediction

The location of a protein within the cell is frequently linked to its function. For example, proteins spanning membrane regions and containing signal peptides

and cleavage sites are commonly associated with the cell secretome (secretory proteins) and function as transport and signaling proteins [Möller et al., 2001]. DeepTMHMM [Hallgren et al., 2022], TargetP v2.0 [Armenteros et al., 2019], SignalP v5.0 [Almagro Armenteros et al., 2019], and Phobius [Käll et al., 2004] were used to identify such proteins amongst the HPs [accessed on January 12, 2023].

2.4.2.3 Analysis of protein physicochemical properties

General physical and chemical features of HPs were computed using the ProtParam tool on the ExPASy web server [Gasteiger et al., 2005a][accessed on January 12, 2023]. The output included a diverse range of properties, including the proteins’ molecular weight (MW), theoretical isoelectric point (pI), amino acid composition, atomic composition, extinction coefficient, instability index, aliphatic index, estimated half-life, and the GRAVY score. An in-house Python script using Jupyter Lab v1.2.6 [Van Rossum and Drake, 2000, Silaparasetty, 2020] was created to enable batch processing of all HPs. The script used a remote command-line search of the HPs through the ProtParam tool.

2.4.2.4 KEGG and Gene Ontology analysis

KEGG v104 [Kanehisa et al., 2016] [accessed on January 5 2023] was used to identify the primary pathways the HPs were involved in, using the KofamKOALA (KEGG Orthology And Links Annotation) analysis tool, which assigns K-numbers to sequences using HMMSEARCH against the customized KEGG Orthologs HMM database. Significant scores were highlighted in the output with an asterisk.

2.4.3 Characterization of five case studies

2.4.3.1 MSA, phylogenetic tree calculations, and pairwise sequence identity

MSA was done on HPs with known/characterized homologs using the Multiple Sequence Comparison by Log-Expectation (MUSCLE) algorithm [Edgar et al., 2004] in the MEGA X v10 software [Kumar et al., 2018]. Batch identification of the best three maximum likelihood tree models and phylogenetic tree calculations were

done in MEGAX using its command-line tool, MEGA-CC [Kumar et al., 2012]. Nine runs were performed for each set of aligned proteins, yielding two outputs per run: the consensus and test neighbor-joining trees at 1000 bootstrap replicates. The generated trees were visualized in Dendroscope v3.7.4 [Huson et al., 2007]. A comparison was made between the calculated and the consensus trees by checking the branching and bootstrap support values. Trees that did not match the consensus were first considered "failed" runs and realignment was done to improve the phylogenetic calculation. The top 5 proteins with the best trees were used for subsequent annotations as case studies (Tree quality was classified as those with bootstrap values greater than 50%). These were VICG00012, VICG01314, VICG01349, VICG01687, and VICG01723. These five HPs from diverse functional groups were selected as case studies for further annotation. Moreover, an all versus all pairwise sequence identity calculation was performed for each protein sequence set using an in-house Python script and a heatmap generated in Jupyter Notebook, as described in [Faya et al., 2015]. To generate the heatmaps presented here, a phylogenetic analysis of the sequence data was first conducted using the Maximum Likelihood method, resulting in a phylogenetic tree illustrating the evolutionary relationships among the sequences. From this tree, pairwise distances or similarities between the sequences were calculated, determining measures such as nucleotide substitutions and genetic distances. This information was compiled into a distance matrix, where each cell represented the similarity or distance between pairs of sequences, with values converted to a percentage scale ranging from 0% (no identity) to 100% (identical sequences). Utilizing JupyterLab, this distance matrix was then visualized as heatmaps, with a color gradient indicating the degree of similarity. Sequence identifiers were labeled on the axes, and a color legend was included to clarify the scale of similarity, providing a comprehensive view of the sequence relationships derived from the phylogenetic analysis.

2.4.3.2 PPI network analysis

PPI networks were analyzed using STRING v11 [Szklarczyk et al., 2019, 2021]. STRING, a computational PPI prediction database, integrates data from organisms and known interactions stored in primary databases, predicting the structural

and functional associations between interacting proteins. In PPI networks analyzed using the STRING online platform, the p-value serves as a critical metric for assessing the statistical significance of observed interactions. A p-value significantly less than the conventional threshold of 0.05 indicates that the interactions are statistically robust and unlikely to be the result of random chance. This low p-value provides strong evidence for the genuine biological relevance of the interactions, thereby enhancing the reliability of the detected associations within the network and supporting the validity of the interaction data. The HP interactions analysis was performed by specifying the microorganism as *V. corneae* and searching the uncharacterized proteins by sequence using default parameters (full STRING network type and medium confidence score = 0.400). The association networks displayed in this study were limited to interaction evidence based on laboratory experiments.

2.4.3.3 Functional annotation: functional domains prediction, motif analysis and homology modeling

The online NCBI Batch CD Search identified HP conserved functional domains [data retrieved in December 2022]. The search was done against the CDD v3.19 [Lu and Salzberg, 2020, Marchler-Bauer et al., 2017] (E-value = 10^{-4}) with filtered low-complexity regions and composition-corrected scoring). This was compared to structural and functional domains identified using the ExPASy ScanProsite program [de Castro et al., 2006]. The search was restricted to the Fungi group. Conserved motifs were identified using MEME [Bailey et al., 2006] and Motif Alignment and Search Tool (MAST) [Bailey and Gribskov, 1998], which are part of the motif discovery MEME Suite online platform v5.3.3 [Bailey et al., 2015]. MEME and MAST command-line versions were used. A maximum of 15 motifs with motif lengths ranging from 3–20 residues [Ross et al., 2017], a range of 2–10 sites per motif, and an E-value of less than 0.05 were identified in MEME. MAST was used to analyze the significance of these motifs using p-values by sorting the sequences in descending order (E-value less or equal to 10; position p-value less or equal to 10^{-5}). Duplicate motifs were removed by rerunning MAST. Selected motifs were mapped onto the respective protein MSAs and 3D models using Jalview

v2.11.1.4 [Waterhouse et al., 2009] and Pymol [Yuan et al., 2017]. Conserved motifs heatmaps were generated using a custom Python script by calculating the total number of sites per the total number of subgroup sequences as previously described [Ross et al., 2017].

Template identification was done using HHSearch on PRIMO [Hatherley et al., 2016] and HHPred [Söding et al., 2005] to identify suitable structural templates for homology modeling. This process is crucial for predicting the three-dimensional structure of a protein by finding homologous structures with known configurations, which can guide the accurate modeling of the structure of the target protein structure. The best template was selected based on the percent sequence identity, high coverage, and high resolution (2-3.80Å), and alignment was done using TCOFFEE. Moreover, pairwise secondary structure alignment of the query and template were analyzed to assess the quality of the alignment and confirm that the selected template was the best for subsequent homology modeling. MODELLER v9.23 [Zimmermann et al., 2018] calculated a hundred models [Söding et al., 2005] using the very slow refinement mode. The top three models with the lowest z-DOPE score were selected for validation. The best models were selected based on quality assessments using ProSA [Wiederstein and Sippl, 2007], PROCHECK [Pontius et al., 1996], QMEAN [Waterhouse et al., 2018], and Verify3D [Luthy et al., 1997]. PyMOL [Yuan et al., 2017] was used to map highly conserved motifs onto the 3D model. Only motifs within the plane of view were labeled on each predicted structure. Motif numbering was based on output labels from the motif-discovery MEME Suite runs.

Table 2.1: A summary of the selected templates used to model each HP

HP ID	Template ID	Sequence Identity (%)	Coverage (%)	Resolution (Å)
VICG00012	6QM7	25	2-231 (98%)	2.80
VICG01314	6AZ3	60	2-377 (87.8%)	2.50
	1CF7	19	27-100 (35%)	2.60
VICG01349	5TUU	24	102-187 (41%)	2.25
VICG01687	2WJY_A	43	2-565 (98%)	2.50
	4BUJ	36	62-900 (%)	3.70
	5MC6	38	64-900 (%)	3.80
VICG01723	4A4Z	38	63-900 (%)	2.40
	6IEH	44	38-900 (%)	2.89
	6BB8	45	58-900 (%)	3.49

2.5 Results and Discussion

The study was conducted in two phases, as shown in Figure 2.1. The initial phase, PHASE I, primarily focused on identifying, retrieving, and curating unique HPs from various online resources through rigorous homology searches. The second phase (PHASE II) involved an in-depth analysis of the selected proteins. This involved the classification of HPs into protein families, analysis of their probable roles in metabolic pathways, characterization of their physicochemical properties, homology modeling, prediction of transmembrane localization and signal peptides, and PPI network analysis. Altogether, this provided substantial evidence of the function and structure of the HPs.

2.5.1 Phase I

2.5.1.1 Sequence retrieval

The proteome of *V. corneae* retrieved from NCBI [accessed on December 8, 2022] consists of 2,237 unique protein sequences [Broad Institute, 2011]. A staggering 2,034 of these proteins (90.97%) were classified as HPs, while only 203 (9.1%) were identified as fully characterized proteins. Among the HPs, 207 were labeled as fragmented/partial protein sequences in the retrieved sequence file and were discarded from the subsequent annotations. The remaining 1,827 were marked as complete protein sequences. Furthermore, the retrieved sequences were subjected to a batch UniProt search to identify the availability of annotation data [accessed in December 2022].

2.5.1.2 Sequence homology search: Distribution and identification of HP homologs

Out of the 1,827 HPs, 627 (34.31%) did not have any homologous sequences in the NCBI database, even after conducting a BLASTp homology search coupled with PSI-BLAST. Sequence similarity is a fundamental approach for predicting protein function, relying on the assumption that homologous proteins often share similar

biological roles. In this process, uncharacterized proteins are compared to proteins with known functions within curated databases. When an uncharacterized protein exhibits significant sequence homology to a well-annotated protein, functional inference can be made with reasonable confidence. However, when an HP shows sequence similarity exclusively to other uncharacterized proteins, functional prediction becomes unreliable. This is due to the absence of functional information in the comparative dataset, leading to a lack of biologically meaningful context for accurate annotation. Given the importance of sequence similarity in protein function inference, any HP that was homologous solely to other uncharacterized proteins was excluded from the dataset. This filtering step was essential to reduce noise and ensure that only proteins with the potential for reliable functional annotation were retained. As a result, 845 sequences, representing 46.25% of the total HPs, were removed from the analysis to maintain the integrity and robustness of the functional predictions derived from this study. The following annotation steps were then conducted on 355 (19.43%) HPs that were hereafter retained. A BLASTp search was conducted for these 355 proteins, and several homologs from different microsporidia species were identified. The top 10 BLASTp hits for each protein were selected to confirm the true homology of the HPs. True homologs were identified as hits from sequences where the HP appeared in the top three with high coverage and percent identity (Figure 2.2). Out of this search, 162 HPs were retained for further downstream in silico characterization processes.

2.5.2 Phase II

2.5.2.1 HPs in *V. corneae* are mostly involved in regulatory processes and the transportation of molecules

The microsporidia proteome has been reported to contain several copies of proteins performing similar functions. Using the programs highlighted in the List of Tools, these proteins were classified into different protein families involved in various functions. This classification was primarily based on the Pfam annotation program [Mistry et al., 2021]. Of 162 HPs, 143 (88.27%) were classified, while 19 HPs did

not belong to any classes. The most significant proportion of classified HPs were binding proteins ($n = 60$, 37%), followed by enzymes ($n = 38$), proteins involved in cellular processes ($n = 13$), and transport proteins ($n = 8$) (Figure 2.3a). Binding proteins were further classified based on the molecule they bind to, including nucleic acids, proteins, ATP, and other substances (Figure 2.3b).

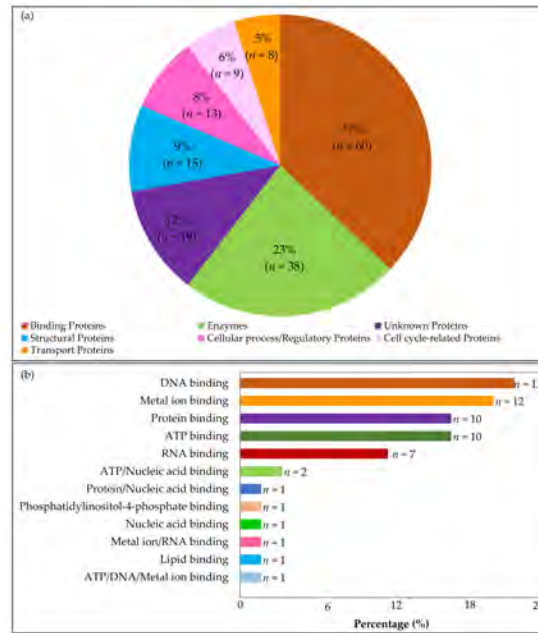


Figure 2.3: **(a)** A pie-chart representation of the different protein family classes identified in *V. corneae* HPsA total of 162 HPs classified into respective functional groups based on functional annotations using several online tools. The number of HPs in each group is illustrated by the number following each group name and their respective percentage in brackets. The chart shows that the bulk of the proteins (37%) are involved in binding to different molecules (nucleic acids, proteins, ATP), followed by enzymes (23%), structural proteins (9%), regulatory proteins (8%), cell-cycle proteins (6%) and transport proteins (5%). Cellular and regulatory proteins include translation, transcription, and replication. The unknown proteins highlighted include those that did not have significant hits across the tools used. **(b)** A bar graph representation of the proportion of binding proteins. Nucleic acid-binding proteins, ATP, and protein-binding proteins. The numbers next to the labels indicate the total number of HPs in each group out of 60 binding proteins, with percentages. The order of each section is sorted by size, with the largest proportion of the binding HPs binding to DNA (13.22%) followed by metal ions (12.2%).

These groups were further grouped into distinct sub-classes (Table [2.5.2.1](#)).

Table 2.2: A summary of the functional categories of the annotated HPs.

Predicted Function		HPs (NCBI Accession numbers)
1. Enzymatic Activity	Hydrolase	XP_007603465.1, XP_007603484.1, XP_007603568.1, XP_007603726.1, XP_007603893.1, XP_007604190.1, XP_007604316.1, XP_007604934.1, XP_007605007.1, XP_007605048.1, XP_007605091.1, XP_007605128.1, XP_007605132.1, XP_007605168.1, XP_007605243.1, XP_007605249.1, XP_007605468.1
	Phosphatase	XP_007605069.1, XP_007605421.1
	Kinase	XP_007603954.1, XP_007604292.1
	Transferase	XP_007603549.1, XP_007603706.1, XP_007604045.1, XP_007604313.1, XP_007604512.1, XP_007604572.1, XP_007604786.1, XP_007605348.1, XP_007605513.1
	Redoxin	XP_007603620.1, XP_007604320.1
	Reductase	XP_007604630.1

Continued on next page

Table 2.2: A summary of the functional categories of the annotated HPs. (Continued)

Predicted Function		HPs (NCBI Accession numbers)
2. Binding Proteins	DNA Binding	XP_007604222.1, XP_007603555.1, XP_007605148.1, XP_007603976.1, XP_007604702.1, XP_007604711.1, XP_007605014.1, XP_007605563.1, XP_007604217.1, XP_007605600.1, XP_007604795.1, XP_007604991.1, XP_007605236.1
	RNA Binding	XP_007603502.1, XP_007603756.1, XP_007604025.1, XP_007603684.1, XP_007604059.1, XP_007603503.1, XP_007605101.1
	Protein Binding	XP_007603813.1, XP_007604439.1, XP_007604066.1, XP_007604530.1, XP_007604889.1, XP_007605221.1, XP_007605241.1, XP_007605535.1, XP_007604255.1, XP_007604458.1, XP_007604339.1,

Continued on next page

Table 2.2: A summary of the functional categories of the annotated HPs. (Continued)

Predicted Function		HPs (NCBI Accession numbers)
3. Other Proteins	ATP Binding	XP_007605468.1, XP_007605243.1, XP_007603484.1, XP_007604934.1, XP_007605249.1, XP_007605007.1, XP_007605091.1, XP_007604805.1, XP_007605346.1, XP_007603899.1, XP_007604183.1, XP_007604190.1, XP_007604236.1, XP_007605168.1
	Metal ion Binding	XP_007603620.1, XP_007604027.1, XP_007605126.1, XP_007605437.1, XP_007603868.1, XP_007604818.1, XP_007604533.1, XP_007603568.1, XP_007603571.1, XP_007603586.1, XP_007603651.1, XP_007604813.1, XP_007605092.1
	Lipid Binding	XP_007603764.1
	Transport	XP_007605235.1, XP_007605292.1, XP_007603481.1, XP_007605504.1, XP_007603688.1, XP_007603592.1, XP_007603628.1, XP_007605468.1

Continued on next page

Table 2.2: A summary of the functional categories of the annotated HPs. (Continued)

Predicted Function		HPs (NCBI Accession numbers)
	Structural	XP_007604846.1, XP_007605182.1, XP_007605382.1, XP_007605476.1, XP_007603578.1, XP_007604054.1, XP_007604133.1, XP_007604178.1, XP_007604216.1, XP_007604260.1, XP_007604760.1, XP_007604961.1, XP_007605096.1, XP_007605153.1, XP_007605458.1
	Cellular/ Regulatory	XP_007603642.1, XP_007604801.1, XP_007604849.1, XP_007603534.1, XP_007604051.1, XP_007603959.1, XP_007604109.1, XP_007604802.1, XP_007603650.1, XP_007604335.1, XP_007604344.1, XP_007603773.1, XP_007603728.1
	Cell-cycle Related Proteins	XP_007604773.1, XP_007604450.1, XP_007604487.1, XP_007604653.1, XP_007603902.1, XP_007603465.1, XP_007603726.1, XP_007603972.1, XP_007604824.1

2.5.2.2 Binding proteins

Based on the molecule they bind to, sixty binding proteins were predicted and classified into distinct groups, including nucleic acid-binding, protein-binding, metal-/metal-ions-binding, ATP-binding, or binding to other molecules (Figure 2.3b). Thirteen HPs were identified as DNA-binding, and eight were RNA-binding. Among them, four HPs (VICG01256, VICG01265, VICG01568, and VICG02118) containing the homeobox domain were involved in DNA-binding and transcription regulation, and one histone-domain-containing DNA-binding protein (VICG00524) was involved in regulating protein heterodimerization activity.

Eleven HPs were classified as protein-binding and characterized by containing WD-repeats domains, PFU domains, or TPR domains. A set of ubiquitin fusion degradation proteins, functioning in post-translational modification, protein turnover, and chaperones, was identified. The XP_007604246.1 protein belonged to the ubiquitin conjugation factor E4 family and was involved in ubiquitin-ubiquitin ligase activity [Koegl et al., 1999]. Thirteen HPs were found to be metal-binding and involved in various cellular processes during DNA replication. Ten contained the zinc finger domain, which can bind to other molecules such as DNA and proteins and enhance their stability and structural integrity in different cellular environments [Iuchi, 2001]. XP_007605437.1, a metal-ion-binding protein containing the HORMA domain, was predicted to be involved in DNA repair.

2.5.2.3 Enzymes

Enzymes are crucial for metabolic processes, and microsporidia reportedly retain many enzymes they secrete into host cells to manage host metabolism [Senderskiy et al., 2014]. About 23% ($n = 38$) of the predicted HPs were identified as enzymes, grouped into several categories, including transferases, kinases, hydrolases, reductases, phosphatases, and proteinases.

Ten of the HPs were transferases, which play an important role in oxidative stress [Pan et al., 2018]. For example, XP_007603549.1 and XP_007604313.1 were classified as S-adenosyl-L-methionine-dependent methyltransferase and histone acetyltransferase, respectively. The former detoxifies cytotoxicity and regulates RNA

methylation, while the latter catalyzes the transcription process [Pan et al., 2018]. Two HPs were predicted to be kinases, which are receptor proteins involved in signal transduction and cell cycle regulation [Miranda-Saavedra et al., 2007, Li et al., 2014]. XP_007603954.1 and XP_007604292.1 were classified as ribokinase-like superfamily and TNase-like domain-containing membranous protein, respectively. XP_007603954.1 was identified as a cytoplasmic protein based on transmembrane localization prediction, contains the NAD(P)HX binding site involved in ADP-dependent NAD(P)H-hydrate dehydratase activity and ATP binding. XP_007604292.1 was characterized as a membranous protein containing the staphylococcal nuclease (SNase) homologs with an open beta-barrel (OB-fold) that mediates DNA-binding.

Hydrolases, crucial for host tissue invasion and evasion of host defense mechanisms [Pan et al., 2018], were predicted to be present in six HPs, with four of them classified as P-loop-containing nucleoside triphosphate hydrolases. These were: XP_007603893.1, XP_007605048.1, XP_007605132.1, and XP_007604316.1.

One HP, XP_007604630.1, was classified as dihydrofolate reductase (DHFR), which regulates levels of folate coenzymes. This HP contained a DHFR domain linked to glycine and tetrahydrofolate biosynthetic processes [Bertacine Dias et al., 2018].

Finally, one HP, XP_007605421.1, was grouped as alkaline phosphatase. It transfers phosphorus-containing groups and has eight transmembrane helices. Microbes secreting phosphatase deplete the host's phosphate levels, one of the key factors that increase the virulence of disease-causing microorganisms [Gupta et al., 2017].

2.5.2.4 Transport proteins

Transport proteins, comprising carriers, transporters, receptors, and signal transduction proteins, participate in the survival of microorganisms by facilitating the movement of useful nutrients and metabolic waste into and out of their cells [Tamim El Jarkass and Reinke, 2020]. Microsporidia genomes have abundant proteins to support their acquisition of resources from their hosts [Tamim El Jarkass and Reinke, 2020]. Several carriers, transporters, receptors, and signal trans-

duction proteins were identified among the *V. corneae* HPs. Four ATP-Binding Cassettes (ABC) transporter proteins were identified: XP_007603484.1, XP_007604934.1, XP_007603628.1, and XP_007605468.1. These ABC proteins regulate the entry and exit of various substrates in the cell, aiding in the attachment of microbes onto the host cell surface [Schneider and Hunke, 1998, Higgins, 2001, He et al., 2019]. These proteins contained a transmembrane domain at the N-terminal involved in pore formation in the inner membrane and a nucleoside triphosphate hydrolase domain that facilitates ATP binding and energy generation [Schneider and Hunke, 1998].

Signal transduction is an essential biochemical process by which cells communicate with extracellular messenger molecules to regulate vital metabolic pathways and control the growth of intracellular microorganisms. The microsporidia proteome is highly enriched with protein domains involved in PPIs, such as signal transduction and WD40 domains [James et al., 2013]. Five signal transduction proteins were identified among the HPs. These included: XP_007603813.1, XP_007604255.1, XP_007604458.1, XP_007604530.1, and XP_007605241. These proteins contain the signal transduction domain and seven repeats of the WD40 domain, which facilitates the PPI system [Xu and Min, 2011, Jain and Pandey, 2018]. These protein domains are essential for the obligatory intracellular behavior of microsporidia, as they allow the microorganisms to interact with the host cells and acquire essential nutrients.

Carrier proteins are another important class of host-cell-interacting molecules found amongst the HPs. These proteins facilitate the importation of nutrients to intracellular parasites. They are characterized by a ferredoxin-containing receptor-binding domain that takes control of the host's iron uptake pathway [Pan et al., 2018]. This analysis revealed the presence of a carrier protein (XP_007603620.1) containing the iron-sulfur (Fe-S) cluster-binding domain..

Receptor proteins are similarly important in cell signaling by binding to signaling molecules and ligands to regulate cellular response, thereby promoting intracellular trafficking of nutrients [Desjardins et al., 2015]. Two receptor proteins were annotated from the HPs: XP_007605437.1 and XP_007605458.1. The HP XP_007605437.1 was further characterized as containing leucine-rich repeats. XP_

007605458.1 was grouped in the PHD finger superfamily containing a nuclear-receptor-binding SET domain, suggesting a possible role in regulating the host cell's gene expression.

2.5.2.5 Structural proteins

Structural proteins are essential in the survival and infectivity of microorganisms by regulating molecular influx from the surrounding host environment and providing a protective protein surface layer against extracellular factors [Kwan et al., 2006]. In microsporidia, proteins such as spore wall proteins also function to adhere to host cells during infections [Yang et al., 2015]. Sixteen HPs were predicted to be structural proteins in this study (Table 2.5.2.1).

The HP XP_007604178.1 was classified as a microsporidian spore wall superfamily protein (MICSWaP). This protein contains a signal peptide at the N-terminal and a non-cytoplasmic domain. The spore wall protein is a major component of the spore wall. It provides a protective barrier against the host immune system, allowing the microsporidia to establish their intracellular niche [Yang et al., 2018]. Identifying this protein provides insights into the molecular mechanisms of the host-microsporidia interaction and may be a potential therapeutic target for microsporidia infections.

The HP XP_007604054.1 was classified as an actin-related protein, a structural constituent of the ribosome linked to the protein's 3D structure. This protein is also associated with the nucleotide-binding domain of the sugar kinase/HSP70/actin superfamily. Actin is involved in filament formation, a major cytoskeleton component that interacts with myosin during cell motility [Kabsch and Vandekerckhove, 1992, Mahadevan et al., 2011]. The presence of actin-related proteins in microsporidia suggests a potential role in the modulation of the host cell's cytoskeleton during infection.

2.5.2.6 Cellular/regulatory proteins

These proteins maintain the microsporidia genome integrity while contributing to regulatory functions and pathogenicity, compensating for the genome reduction observed in these microorganisms [Texier et al., 2010, Peyretailade et al., 2011]. The cellular functions these proteins are involved in include cell-cycle regulation, transcription, translation, and replication. Twenty-seven HPs in the *V. corneae* datasets were classified as cellular/regulatory proteins and included transcription factors and transcription regulatory proteins ($n = 9$), polymerases ($n = 8$), translation proteins ($n = 5$), and ribosomal biogenesis regulatory proteins ($n = 5$).

2.5.3 Transmembrane localization and signal peptide prediction

The function of a protein is often linked to its localization within a cell. Membrane-spanning proteins, for instance, are known to play a role in signaling pathways, and the transportation of nutrients across different biological environments within and outside a cell [Möller et al., 2001]. Therefore, it is crucial to identify the presence of signal peptides and transmembrane helices in HPs. The analysis of subcellular localization of HPs using TMHMM, DeepTMHMM, TargetP, SignalP, and Phobius indicated that HPs were membranous in at least three of these tools. Of the tools used, DeepTMHMM predicted 20 transmembrane proteins, 15 signal peptides, 125 globular, and 4 HPs spanning both the transmembrane and signal peptide regions.

These results were compared to the GRAVY index score of hydropathy. Membrane proteins are typically hydrophobic, and HPs containing transmembrane helices would fall under this category, with GRAVY scores of more than zero. In contrast, globular proteins are localized within the cell and would further be classified as hydrophilic. This analysis displayed the same pattern, with 30 HPs assessed to be hydrophobic, 15 of which were also predicted to be localized on the membrane by different localization prediction tools and 5 containing signal peptides (Appendix Table A1. Out of the 30 hydrophobic HPs, five were predicted to

have signal peptides in both TargetP and SignalP, while 6, 20, and 18 of these HPs with signal peptides were identified with Phobius, TMHMM, and DeepTMHMM, respectively.

The TMHMM tool predicts the number of transmembrane helices in proteins and highlights the expected number of amino acids in the identified helices. If the predicted number is higher than 18, then the probability of the query protein being transmembrane or having a signal peptide is also high. Moreover, transmembrane helices in the N-terminal are more likely to be signal peptides. However, TMHMM cannot identify if the protein is cytoplasmic, unlike other tools.

The Phobius tool predicts if a protein is cytoplasmic or non-cytoplasmic alongside signal peptide prediction. This tool scores the most probable location and orientation of transmembrane helices using the N-best algorithm, similar to the TMHMM tool. Similarly, TargetP and SignalP identify the presence of signal peptides and the regions of the sequence they span. In this survey, TargetP, SignalP, and Phobius predicted 15, 12, and 22 signal peptides, respectively. The combined use of all these tools provides a more comprehensive prediction of membranous and non-membranous proteins.

2.5.4 Physicochemical characterization of HPs

Categorizing proteins based on their physical and chemical properties can help determine their stability and characteristics. This helps plan future laboratory experiments. Analyzing the physicochemical properties of the proteins can provide insights into their potential biochemical functions [Korkuc and Walther, 2015, Moses et al., 2016, Thakur et al., 2020]. In this study, out of the total 164 HPs, 30 of them (18.3%) were hydrophobic, while the remaining 134 (81.7%) were hydrophilic or globular in nature, as illustrated in Figure 2.4a below.

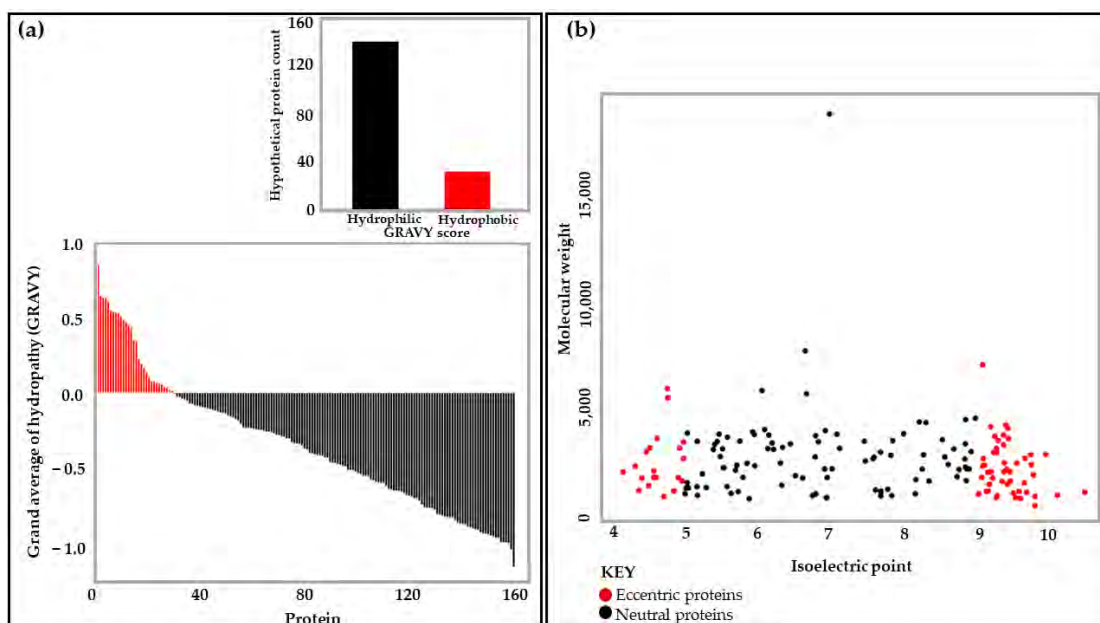


Figure 2.4: Overview of the physicochemical properties of the annotated HPs: (a) The GRAVY scores of less than 0 indicate the prediction of hydrophilic (globular) proteins, while scores larger than 0 are predicted to be hydrophobic (membrane) proteins. A total of 134 (81.7%) HPs were classified as hydrophilic, while 30 (18.3%) HPs were identified as hydrophobic. (b) Isoelectric point vs. molecular weight of HPs using the ProtParam tool. In red are the eccentric proteins with acidic pI (3–5) ($n = 20$; 12.2%) or basic pI (9–12) ($n = 49$; 29.87%), while those in black are proteins with pI range 5–9 ($n = 95$; 57.93%). The majority of HPs fall in this category

The molecular weight and pI provide information on the chemical characteristics of HPs. A significant proportion of HPs were neutral, with pI ranging between 5 and 9 ($n = 95$, 57.93%). In comparison, the number of eccentric HPs (basic or acidic) was lower, with the proportion of essential proteins being 29.87% ($n = 49$). In contrast, that of acidic proteins was 12.2% ($n = 20$) (Figure 2.4b above).

2.5.5 KEGG and Gene Ontology analysis

Identifying protein families provides a multifaceted in silico approach to conclusively annotate undescribed proteins. To confirm the predicted roles of annotated HPs in different metabolic pathways, these proteins were cross-checked against the KEGG pathway database (Appendix Table A2. Of 164 HPs, 55 were accurately annotated, consistent with the protein family classification in Figure 2.3 above. An extensive set of predicted proteins was identified as involved in enzymatic and cell regulatory processes.

2.5.6 Characterization of Representative Proteins

The selection of these five representative HPs (VICG00012, VICG01314, VICG01349, VICG01687 and VICG01723) for detailed structural, PPI, and phylogenetic analyses was based on their potential functional relevance and diversity within the dataset. These proteins were prioritized based on several criteria, including their distinctive sequence features, predicted functional roles, and their ability to represent a range of protein families or classes. This selection allowed for a more focused and in-depth analysis of key proteins that are likely to yield significant insights into the biological processes of interest. By concentrating on a smaller set of representative HPs, the study aimed to achieve a more detailed exploration of protein function, interactions, and evolutionary relationships, which would be less feasible across a broader dataset. This targeted approach ultimately enhances the depth, relevance, and impact of the findings.

2.5.6.1 Case study I: VICG00012

2.5.6.1.1 Identification and analysis of conserved motifs

The HP VICG00012 was identified as belonging to the proteasome alpha subunit (PSA) family based on its homology to other microsporidia species involved in post-translational modification, protein turnover, and chaperones. Motif analysis using MEME identified eight highly conserved motifs across all homologous sequences (Figure 2.5a). Motif 9 was noticeably absent in VICG00012 and its respective *Enterocytozoon hepatopenaei* homolog. *Enterocytozoon hepatopenaei* has been taxonomically identified as a close relative to *V. cornea* [Park and Poulin, 2021]. This could explain the similarity in motif distribution between these two microsporidia

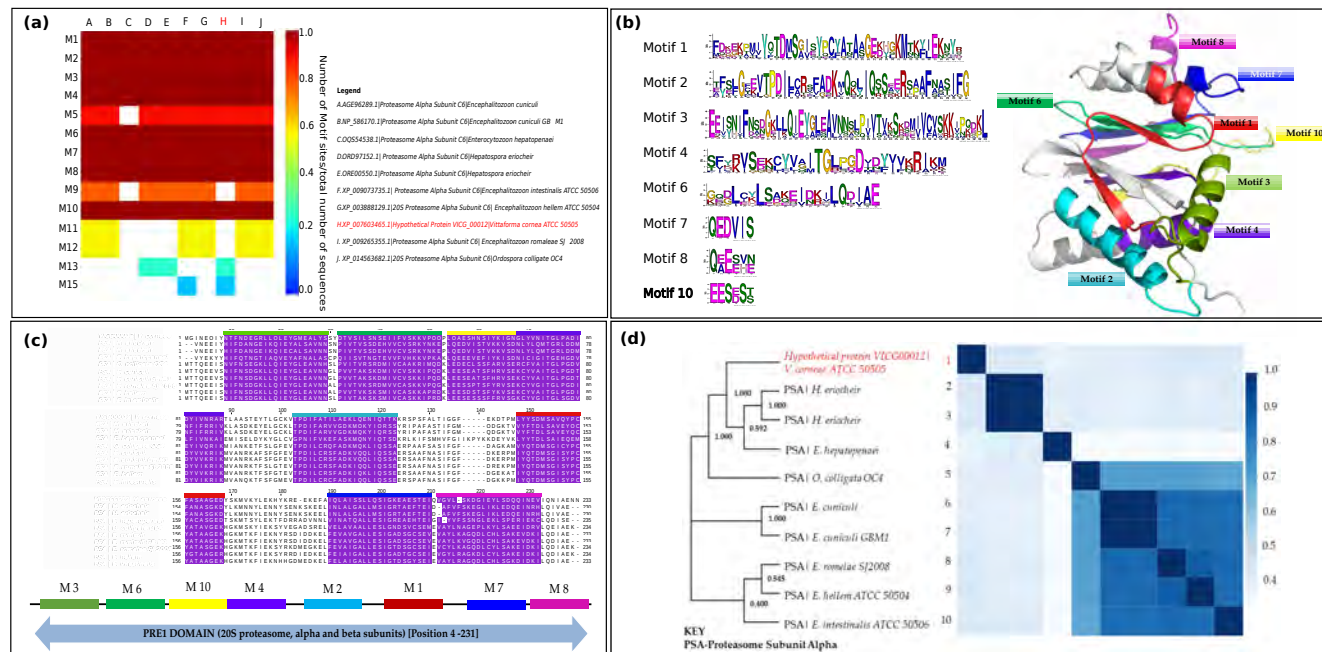


Figure 2.5: Functional and structural annotation of VICG00012. (a) The motif analysis heatmap generated from MEME Suite shows six highly conserved motifs across the PSA homologs, and motif 9 was missing in certain species. (b) Motif mapping to the homology model of the HP. Motifs numbers are preserved based on MEME outputs.(c) Motifs were mapped onto the MSA results of VICG00012 and its homologs, highlighting in purple all motifs conserved throughout the dataset. The conserved domain identified in this sequence belonged to the PRE1 superfamily.(d) Sequence identity heatmap highlighting a pairwise comparison of each of the sequences used in the phylogenetic analysis. A darker color shows a closer identity to respective sequences. Two distinct clusters are observed with higher identity in the *Encephalitozoon* family compared to the clade containing the HP of interest.

2.5.6.1.2 Homology modeling

To design a 3D model of VICG00012, the *Leishmania tarentolae* proteasome 20S subunit protein (PDB ID: 6QM7) was selected as a template for homology modeling. T-COFFEE was used for target-template alignment, and the predicted models with the lowest z-DOPE score were selected. The quality of the modeled structure was assessed using several tools, including Verify3D, QMEAN, PROCHECK, and ProSA-web, which collectively validated the accuracy of the predicted structure. The resulting 3D model of VICG00012 is shown in Figure 2.5b, with eight conserved motifs identified among all species mapped onto the predicted structure using PyMOL. Motifs 13 and 15 were conserved in *V. corneae* and the *Enterocytozoon hepatopenaei* PSA C6, and the motifs were characterized with alpha-helix coils or mapped onto beta sheets in the predicted 3D model.

2.5.6.1.3 MSA and conserved domain analysis

(Figure 2.5c). The HP VICG00012 contains the N-terminal of the proteasome alpha subunit (Figure 2.5c) The PRE1 domain, which is a crucial component of the 20S proteasome structure and is actively involved in post-translational modification, chaperones, and protein turnover [Zwickl et al., 1992, Groll et al., 2000], is present in the HP VICG00012. This domain spans from position 4-231 and represents the N-terminal of the proteasome alpha subunit, as shown in Figure 2.5c.

2.5.6.1.4 Phylogenetic tree analysis and pairwise sequence identity calculations

The phylogenetic analysis of the PSA sequences, including VICG00012, revealed the presence of 3 distinct clusters (Figure 2.5d). VICG00012 PSA homolog was found to be closely associated with PSA from *H. eriocheir*, *E. hepatopenaei*, and *Ordospora colligata* OC4. These findings indicate that this gene may evolve rapidly in microsporidia, aiding their survival in different localization patterns.

The tree analysis results were consistent with the pairwise sequence identity calcu-

lations, revealing that the separated clades had higher sequence identities (Figure 2.5d). The second and third clades exhibited greater sequence similarity than the first group, which included the HP. This similarity could be explained by the fact that all the species in the last two groups infect mammalian hosts [Vossbrinck and Debrunner-Vossbrinck, 2005, Park, 2020].

2.5.6.1.5 PPI network analysis

A novel polar tube protein, a key feature involved in microsporidia spore germination, was identified based on clustering with other essential proteins found in the structure of the microsporidia [Delbac et al., 2001, Polonais et al., 2013]. These studies showed the capacity of using information from neighboring/closely interacting proteins to decipher the function and structure of unknown proteins. The HP VICG00012 (Accession number: XP_007603465.1) has a high degree of homology to proteasome subunit proteins from other microsporidia species and is predicted to function in the proteolysis process involved in cellular catabolic processes. Its homologous superfamily is identified as nucleophile aminohydrolase (Ntn hydrolases), and it is located within the proteasome core complex, containing a threonine nucleophile. The VICG00012 protein was found to have strong interactions with PCI-domain-containing proteins, which include α -helices and a winged-helix domain (Figure 2.6). The p-value for these interactions is 4.66×10^{-15} indicating an extremely high level of statistical significance. This low p-value suggests that the observed interactions are highly unlikely to have occurred by chance, affirming the robustness and reliability of the findings in the VICG00012 protein-protein interaction network analysis. The protein was also closely associated with similar-functioning proteasome subunit proteins L2GLB0, L2GLC6, L2GLK0, and L2GMP0, as indicated by the pink edges in the experimentally determined associations (Figure 2.6).

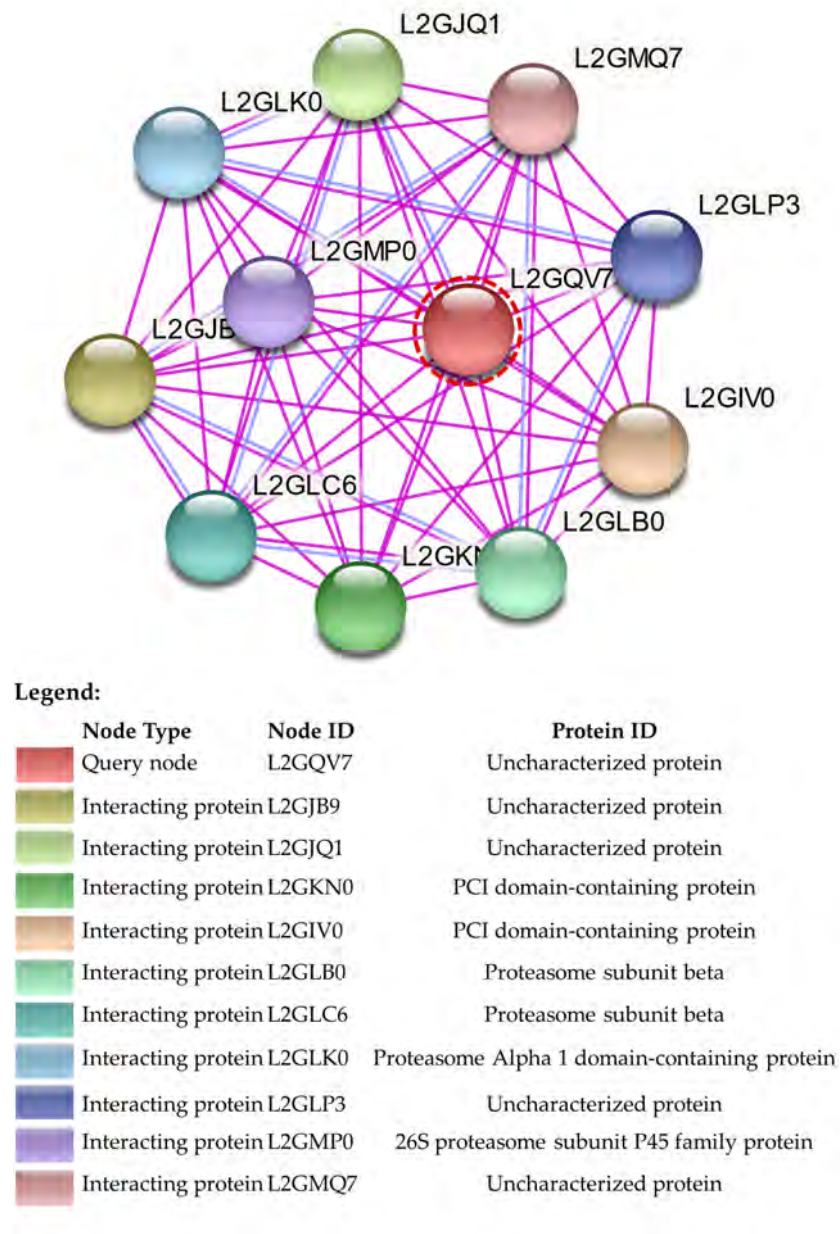


Figure 2.6: A PPI network analysis using the STRING database, with the query HPs highlighted in red (L2GQV7). Each node represents a protein, and a total of 10 interacting proteins (nodes) were observed, with 55 edges indicating both physical and functional protein associations. The color of the edges indicates the type of interaction evidence, with pink edges indicating experimentally-determined associations and purple edges indicating protein homology. The analysis revealed that the HP has close homology to a proteasome alpha 1 domain-containing protein (L2GLK0) and is experimentally linked to similar proteins. The protein-protein interaction (PPI) enrichment p-value is 4.66×10^{-15} , indicating that the clustered proteins in this group are at least partially biologically connected.

2.5.6.2 Case study II: VICG01314

2.5.6.2.1 Identification and analysis of conserved motifs

VICG01314 (NCBI accession number: XP_007604760.1) was identified as a homolog to the ribosomal protein L3, which is involved in protein translation and located in the ribosome. Ribosomal protein L3 forms a part of the large ribosomal subunit and is predicted to play a role in regulating metabolic processes. The HP and its homologs contained nine highly conserved motifs (Figure 2.7a). Motif 15 was present only in *V. corneae* and *Encephalitozoon intestinalis* ATCC 50506 as indicated below.

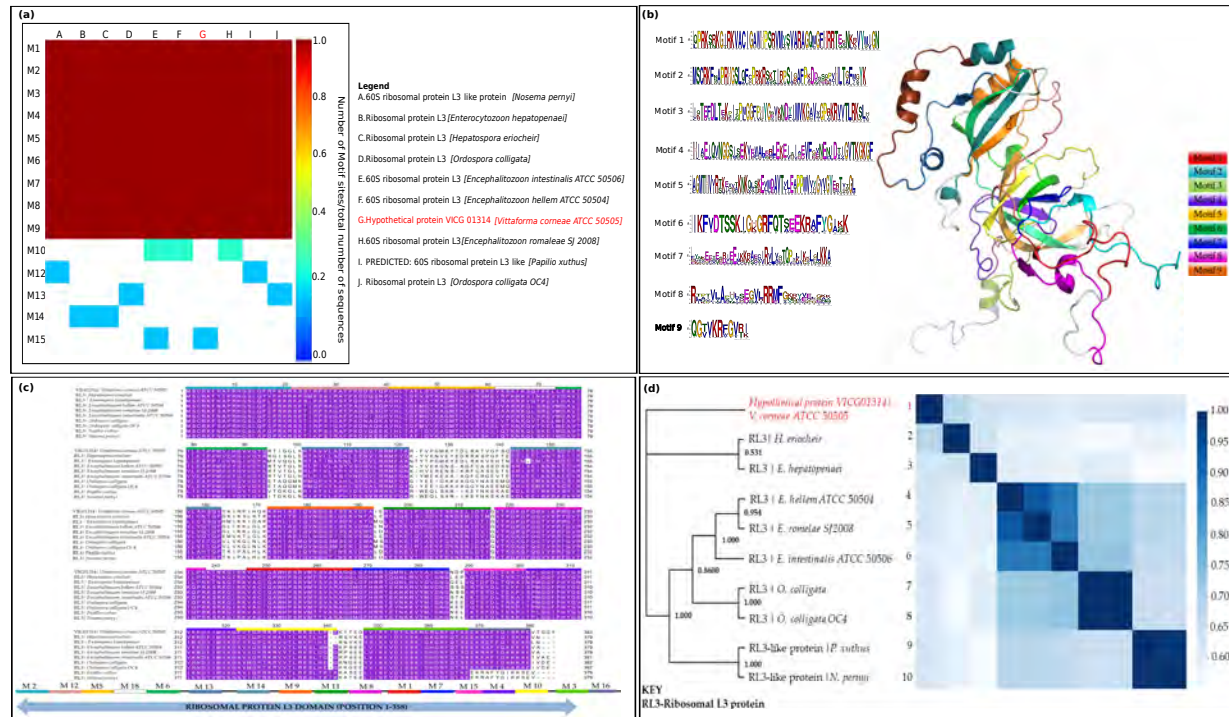


Figure 2.7: Structural and functional analysis of VICG01314. (a) Motif analysis of the HP showing nine conserved regions. (b) Mapping of the motifs onto the 3D structure of the HP (c) MSA illustrating, in purple, several conserved regions in this set of sequences. (d) Phylogenetic analysis of VICG01314 using the Maximum Likelihood method shows that the HP VICG01314 clustered away from the rest of the clades but was closer relative to *H. eriocheir* and *E. hepatopenaei* (The heatmap scale ranges from 0-100%).

2.5.6.2.2 Homology modeling

The 3D model of the HP was structured using the 6AZ3 chain B template, a ribosomal protein L3 from the *Leishmania* parasite (Figure 2.7b).

2.5.6.2.3 MSA and conserved domain analysis

VICG01314 contains the ribosomal protein L3 domain (residue 1-358), which is conserved across all sequences (Figure 2.7c). MSA and motif mapping highlight very few residue-level differences within these sequences.

2.5.6.2.4 Phylogenetic tree analysis and pairwise sequence identity calculations

Phylogenetic analysis shows that VICG01314 clustered away from the rest of the clades but was closer relative to *H. eriocheir* and *E. hepatopenaei* (Figure 2.7d). Pairwise sequence identity calculations determined that the sequences were of close evolutionary distances

2.5.6.2.5 PPI network analysis

The STRING protein association prediction suggests that VICG01314 is closely associated with other regulatory and enzymatic proteins such as translation initiation factor, synthases, kinases, and GTP-binding proteins (Figure 2.8).

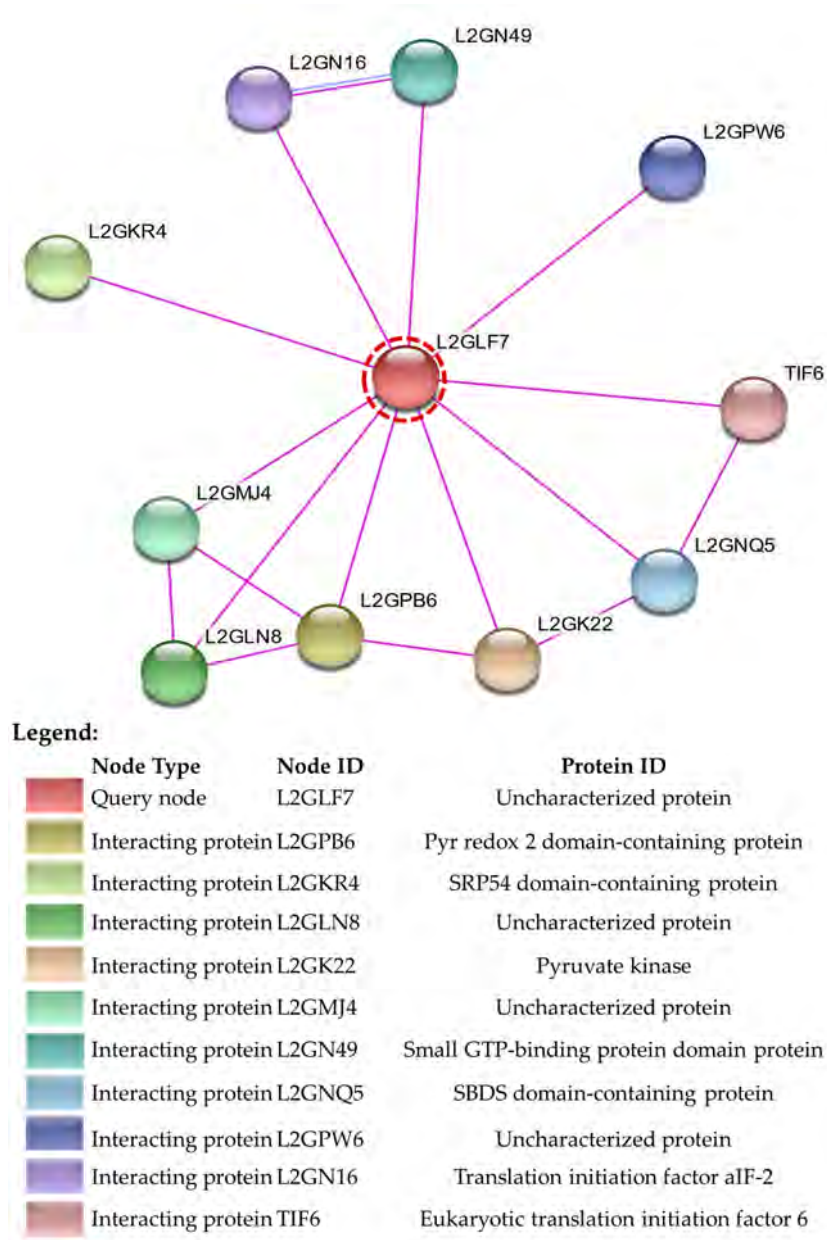


Figure 2.8: Protein association prediction shows VICG01314 to be in close proximity to other regulatory and enzymatic proteins with varying functions such as translation initiation factor, synthases, kinases, and GTP-binding proteins

2.5.6.3 Case study III: VICG01349

2.5.6.3.1 Identification and analysis of conserved motifs

The protein VICG01349 (NCBI accession number: XP_007604795.1) belongs to the DP-DD superfamily based on homology and is classified as the E2F-dimerization-partner transcription factor. Proteins in this group regulate genes responsible for DNA synthesis, cell cycle progression, apoptosis, and proliferation by forming heterodimers with E2F transcription factors [Zheng et al., 1999]. Motif analysis showed four highly conserved motifs between VICG01349 and its counterparts. Motifs 6, 7, 9, and 10 are conserved only in the HP and its homolog from *H. eriocheir* (Figure 2.9a).

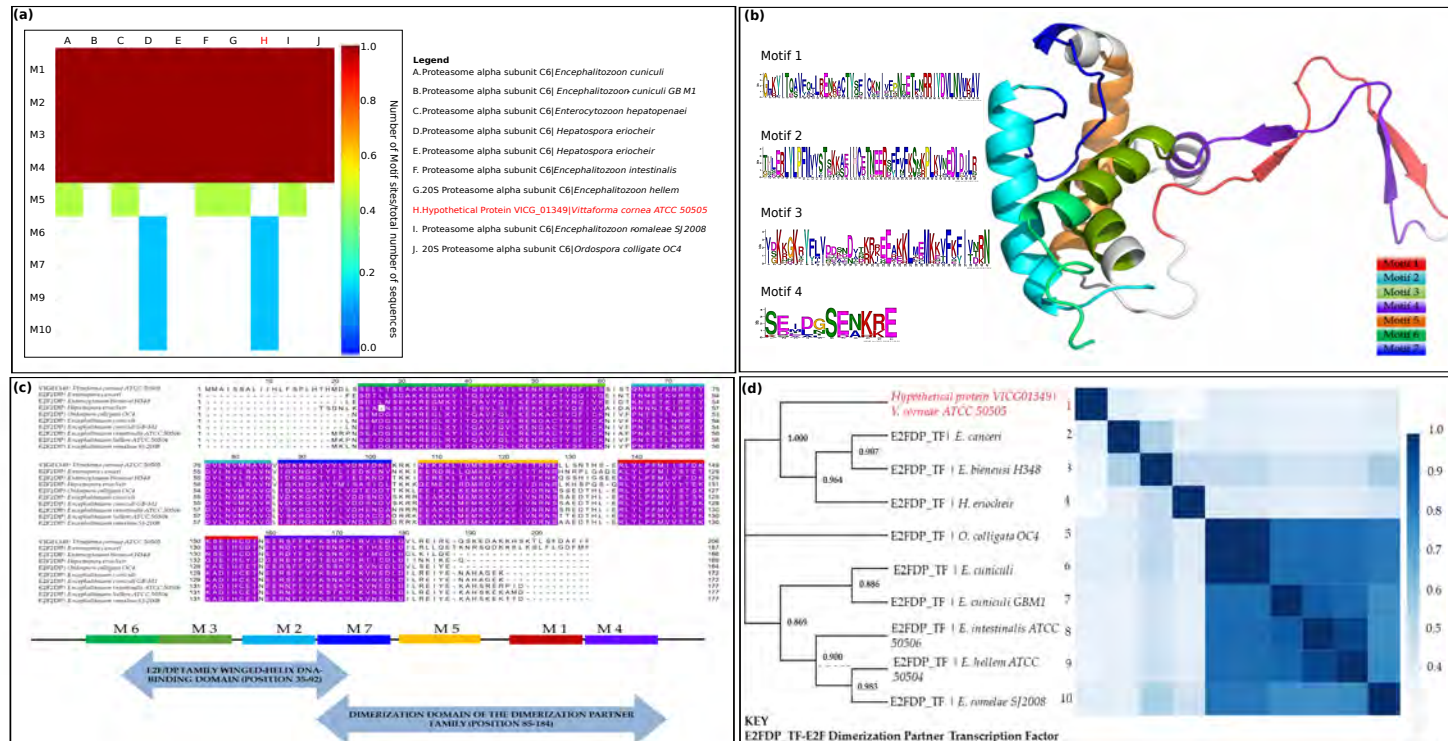


Figure 2.9: Structural and functional analysis of VICG01349. (a) Motif heatmap of VICG1349 showing conservation of 7 motifs across all species. Species and protein IDs are indicated in the legend. (b) The modeled 3D structure of HP VICG01349 with the seven conserved motifs mapped onto the structure. (c) Multiple sequence alignment and motif mapping on VICG01349. Two main domains (the winged-helix DNA-binding domain and the Dimerization domain) are also indicated spanning the region across the sequence as a diagrammatic representation. (d) Tree calculations using the Maximum Likelihood method and sequence identity heatmap of VICG01349 and its homologs illustrate the close relationship between proteins within the same clade (heatmap scale from 0-100%).

2.5.6.3.2 Homology modeling

Figure 2.9b illustrates the 3D structure of the HP as predicted using the transferase protein from *Oryctolagus cuniculus* (PDB ID: 1CF7) and the *Homo sapiens* E2F4-DP1 homolog (PDB ID: 5TUU).

2.5.6.3.3 MSA and conserved domain analysis

VICG01349 is predicted to be located in the cell nucleus and has a signal peptide and non-cytoplasmic domain. Its protein sequence contains a DNA-binding winged-helix motif and a DP domain (Figure 2.9c).

2.5.6.3.4 Phylogenetic tree analysis and pairwise sequence identity calculations

Evolutionary analysis suggests that the protein is less divergent and shows higher identity with both mammalian-specific (*E. bieneusi*) and crustacean-specific microsporidia (*E. canceri*, and *H. eriocheir* (Figure 2.9d).

2.5.6.3.5 PPI network analysis

Key interacting proteins to VICG01349, based on STRING prediction, include proteins in the cyclin family involved in cell-cycle regulation and DNA polymerase and other synthesis proteins involved in a similar role (Figure 2.10).

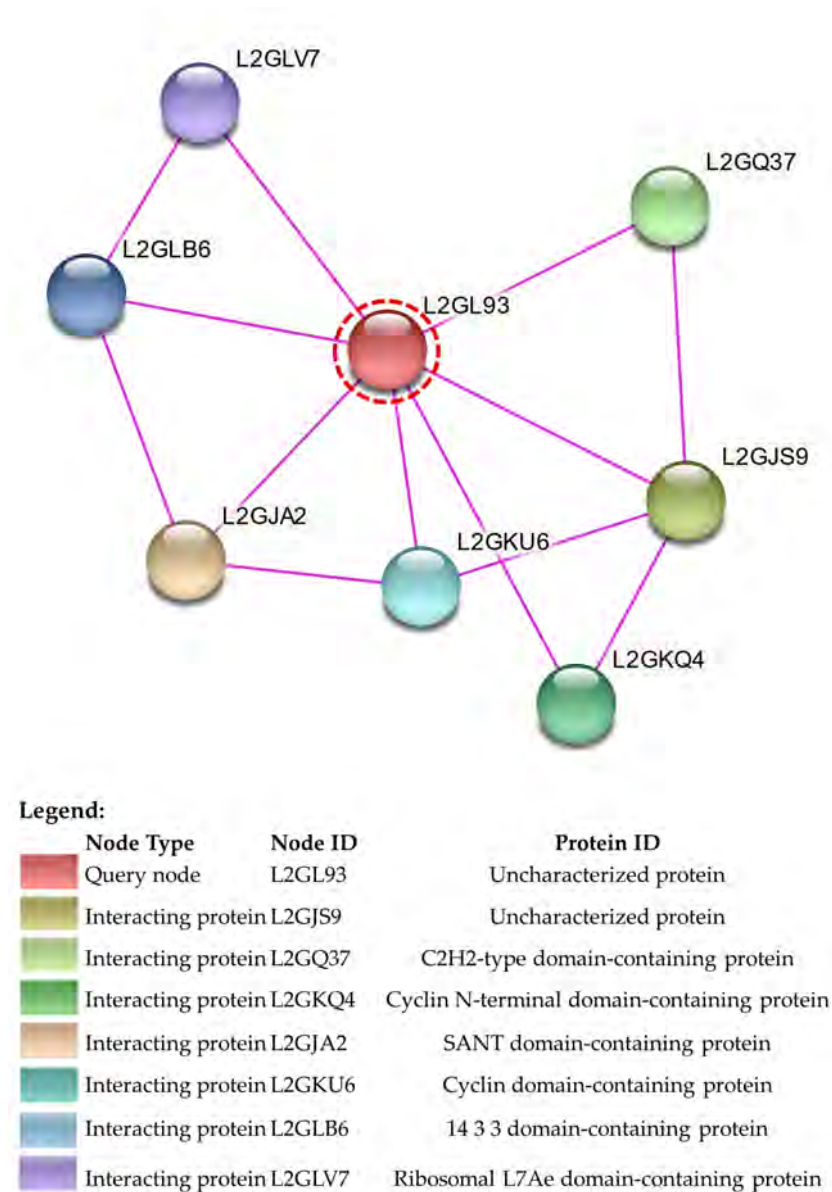


Figure 2.10: PPI analysis of HP VICG01349 using STRING analysis. Close associations with proteins involved in cell cycle processed is indicated.

2.5.6.4 Case study IV: VICG01687

2.5.6.4.1 Identification and analysis of conserved motifs

Helicases can bind to double-stranded DNA and separate them into single-stranded DNA during DNA replication. VICG01687 (NCBI accession number: XP_007605132.1) grouped in the superfamily I DNA-RNA Helicase. Twelve motifs were highly conserved across all homologs of this protein. Motifs 8 and 15 were missing in this HP (Figure [2.11a](#)).

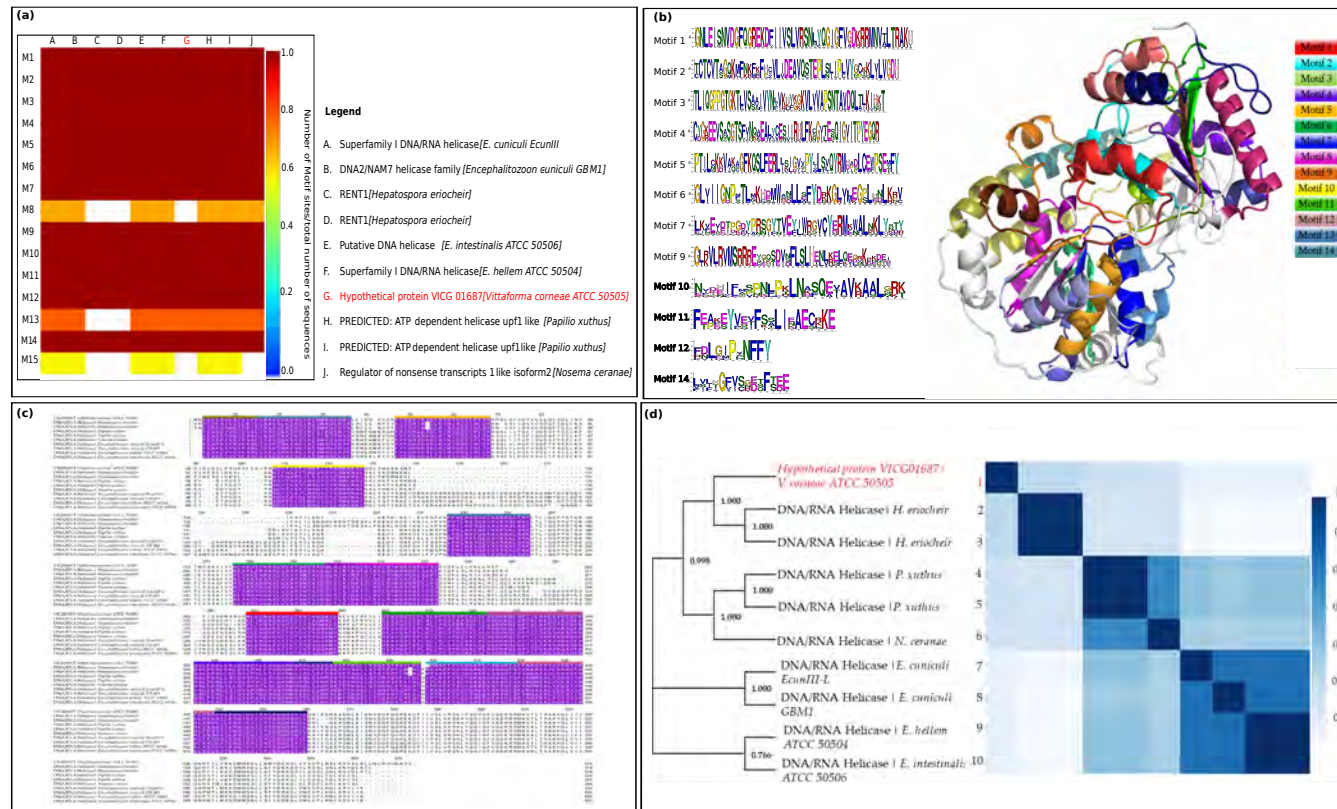


Figure 2.11: (a) Motif analysis heatmap of the protein VICG01687, highlighted in red. A large number of motifs were conserved among these proteins. (b) The 3D model of VICG01687 as predicted using Modeller. Motif mapping is based on the most conserved motif identified using MEME. (c) Multiple sequence alignment and motif mapping of VICG01687. Two main domains are mapped onto the diagrammatic representation of the one-dimensional sequence illustrated below the alignment. (d) Phylogenetic analysis using the Maximum Likelihood method and pairwise identity calculations of VICG01687 and its homologs. Two distinct clusters are observed (Heatmap scale ranging from 0% to 100%).

2.5.6.4.2 Homology modeling

The predicted 3D structure was modeled using the human zinc-finger hydrolase protein (PDB ID: 2WJY), known to regulate nonsense transcripts. Motif mapping is highlighted in Figure 2.11b.

2.5.6.4.3 MSA and conserved domain analysis

Proteins in the helicase superfamily contain a P-loop motif (Figure 2.11c). The HP is characterized by a superfamily 1 C-terminal helicase domain under the Upf1-like helicases family (SF1_C_Upf1) and contains the DEAD domain involved in ATP-dependent RNA and DNA winding [De La Cruz et al., 1999] (Figure 2.11c). Its conserved features include an ATP-binding site and a DNA-binding site involved in transcription regulation [Fuller-Pace, 2006].

2.5.6.4.4 Phylogenetic tree analysis and pairwise sequence identity calculations

Three distinct clusters were identified through studies on phylogeny, and VICG01687 was found to cluster with microorganisms infecting invertebrates, distinctly separated from the microsporidia of mammals. Pairwise sequence identity calculations and the generated heatmap also highlighted this relationship (Figure 2.11d).

2.5.6.4.5 PPI network analysis

VICG01687 was found to interact closely with other binding proteins, including nuclear cap-binding protein, GTP-binding, ribosomal L18e/L15P, hydrolases, and DEAD-box helicase proteins, according to the STRING association network analysis (Figure 2.12). These neighboring proteins are classified within the same superfamily and in proximity to each other, suggesting that they function similarly to VICG01687.

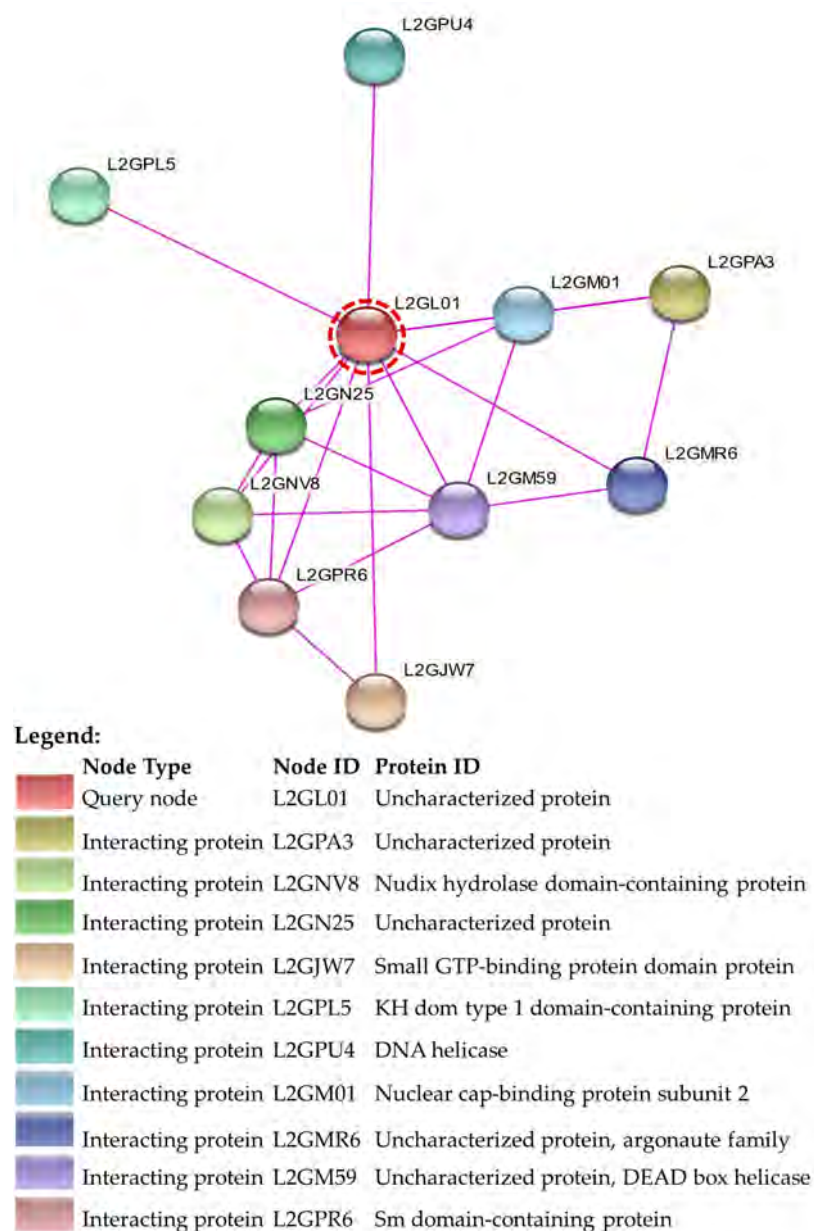


Figure 2.12: STRING analysis of the proteins interacting with VICG01687. The query node indicates the HP VICG01687. Its closest neighbors are proteins from the helicase and argonaute family.

2.5.6.5 Case study V: VICG01723

2.5.6.5.1 Identification and analysis of conserved motifs

A homology search identified VICG01723 (NCBI accession number: XP_007605168.1) as an ATP-Dependent RNA Helicase (SKI2 RNA Helicase). Eleven motifs were

conserved throughout all the microsporidia species analyzed, while motifs 11, 12, and 15 were visibly missing in this HP (Figure [2.13a](#)).

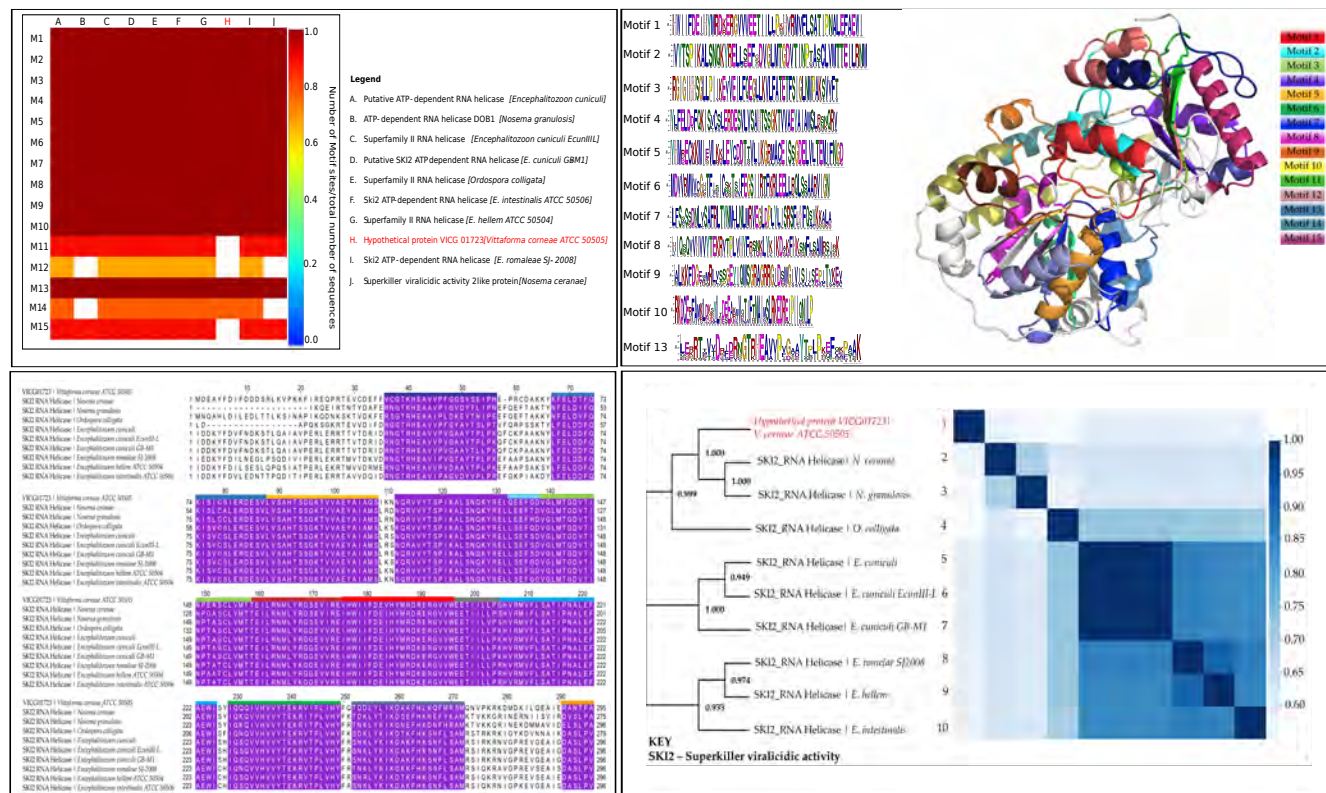


Figure 2.13: (a) Motif analysis heatmap of VICG01723 shows a high number of conserved motifs. (b) The 3D modeled structure of VICG01723 with motif mapping. (c) Multiple sequence alignment of the HP VICG01723 with motif mapping respective to MEME results. Motif mapping is based on the numbering of the identified motifs as determined using MEME. MEME analysis was used to detect conserved sequence motifs across the protein family, which are represented in the alignment. An N-terminal domain (DEAD-BOX HELICASE) and a C-terminal domain are highlighted in the sequence. (d) phylogenetic classification of HP VICG01723 using the Maximum Likelihood tree calculation method and pairwise sequence calculations highlight the sequence identity values among the homologous species (Heatmap scale: 0-100%).

2.5.6.5.2 Homology modeling

VICG01723 protein modeling was done using four templates that covered the length of the protein. The templates used included the antiviral RNA helicase SKI2 protein from *Saccharomyces cerevisiae* S288C (PDB ID: 4BUJ and 5MC6), *Saccharomyces cerevisiae* (PDB ID: 4A4Z), human RNA helicase (PDB ID: 6IEH), and *Neurospora crassa* (PDB ID: 6BB8). The modeling of the VICG01723 protein was carried out using four distinct templates to achieve better coverage across the full length of the protein. Owing to the varying length of these proteins, a single template would not have provided sufficient structural coverage. By utilizing multiple templates, each contributing to specific regions, the overall model was more accurate and reliable, allowing for a more comprehensive representation of the protein structure. This approach ensured that key structural features were adequately captured, which would not have been possible with a single low-coverage template. The identified motifs were mapped onto the 3D structure (Figure 2.13b).

2.5.6.5.3 MSA and conserved domain analysis

The Nop domain is present in pre-RNA ribonucleoproteins and has RNA- and protein-binding surfaces [Fuller-Pace, 2006]. This HP belongs to Superfamily II RNA helicases involved in RNA replication, recombination, and repair. The sequence consists of three distinct domains - the DEAD-box domain, which plays various roles in RNA metabolism; the conserved helicase carboxyl-terminal domain usually found with the DEAD domain; and the DSHCT domain [De La Cruz et al., 1999, Fuller-Pace, 2006] as observed in Figure 2.13c.

2.5.6.5.4 Phylogenetic tree analysis and pairwise sequence identity calculations

The evolutionary relationship analysis demonstrated that these HP clusters are in the same clade as other characterized RNA helicase proteins from microsporidia infecting invertebrates. This HP diverges from similar proteins isolated from mi-

crosporidian species affecting the same host (Figure 2.13d), indicating the versatility of the fast-adapting microsporidia species. This could be due to horizontal gene transfer or gene loss over time, leading to divergence [Keeling and Palmer, 2008, Pombert et al., 2012, Pan et al., 2013].

2.5.6.5.5 PPI network analysis

VICG017223 is linked to other uncharacterized proteins, poly(A) polymerase, the Nop-domain-containing protein, and the WD-repeats region domain-containing protein (Figure 2.14) based on association networks.

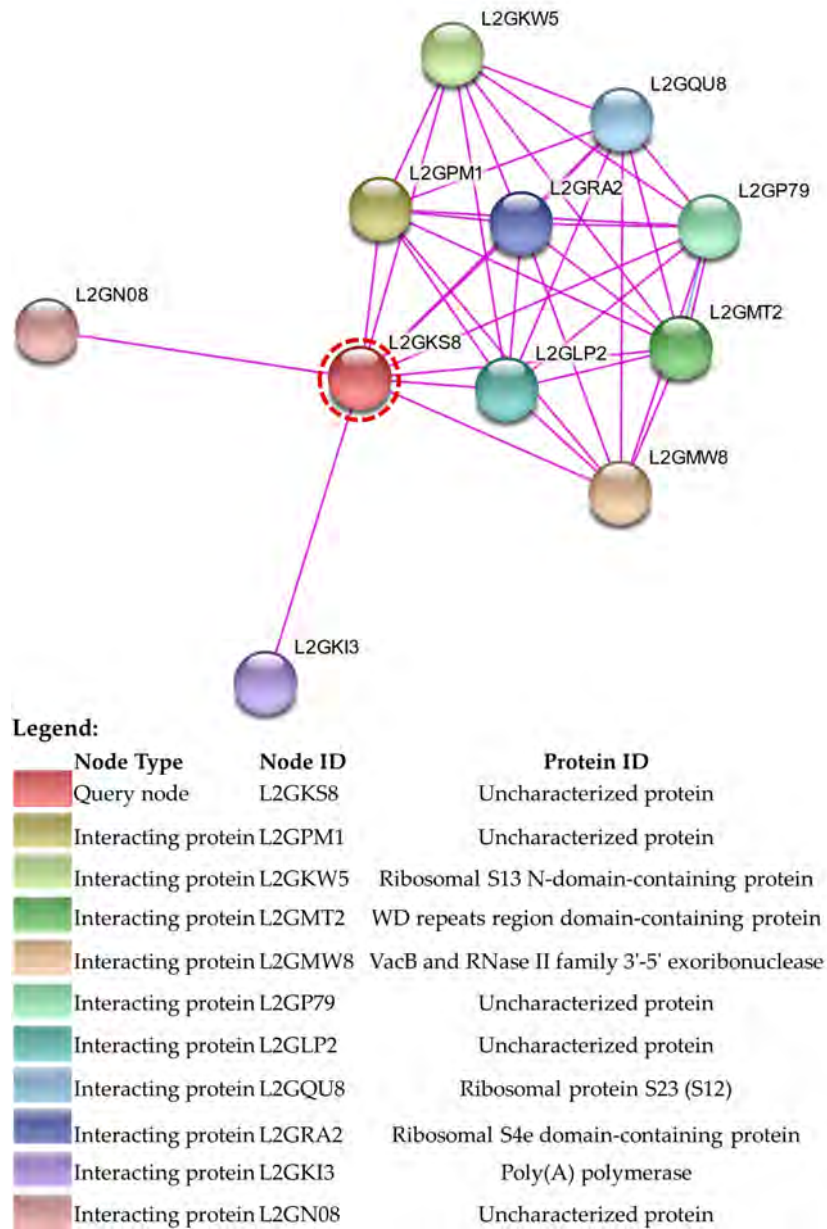


Figure 2.14: PPI network analysis of VICG01723 (query node in red) showing close association with ribosomal proteins.

2.6 Conclusion

Bioinformatics approaches are frequently employed to identify unknown proteins' biochemical and physical properties. With the development of computational tools currently on the fast track, accurate predictions have been possible. The *in silico* identification of significant proteins has improved the processing time of unknown proteins. Studies have highlighted the capacity of these tools to decipher the biological processes in which HPs from different microorganisms are involved. This study uses a multifaceted computational annotation pipeline to characterize the functions and structures of various HPs from *V. corneae*. The HPs annotated in this work, like in most microsporidian species, were found to be involved in regulatory and transport functions, crucial for the survival of these microorganisms due to their compact genomes and severe gene loss. Numerous transporter proteins were identified among the HPs in this microsporidia genome and are predicted to be essential for acquiring host nutrients to support microsporidia development. Identifying these transporters, carriers, receptors, and signal transduction proteins lays the foundation for further studies focusing on host-symbiont molecular interactions and the characterization of this important obligate intracellular organism. The study also highlights conserved regions in these HPs and their relatives, which can be used in the taxonomic classification of the fast-evolving microsporidia species. The chapter also compares several online tools in predicting protein families and functions of these HPs. It demonstrates the capacity of these computational methods to discern the structure and function of unknown proteins. Despite the extensive computational analysis of these proteins, this forms the baseline for future experimental studies to validate these results further. Moreover, the structural information reported in this work can be used in future studies to evaluate interactions of HPs with host proteins. It can be applied in drug discovery and diagnostics.

Chapter 3

De novo Genome Assembly of *Microsporidia* sp. MB isolated from *Anopheles arabiensis* in Kenya

3.1 Chapter Overview

Symbiotic microsporidia are often involved in host protection against other pathogens through immune priming, nutritional competition, and increased development rate- making them effective alternatives to the current vector-control strategies. One such organism, *Microsporidia* sp. MB has been reported to significantly block malaria development in the *Anopheles arabiensis* mosquito and is a promising candidate for malaria control. It is particularly interesting due to its effective mode of transmission from mother to offspring, its presence in multiple *Anopheles* species, its positive effects on the fitness of its host, and its unique life cycle adaptations, making it an intriguing subject for research in mosquito-borne disease control. This study aimed to sequence the genome of *Microsporidia* sp. MB and investigate its genome content to understand better its infection biology, which is an essential element in its future application in field studies. The genome of *Microsporidia* sp. MB was sequenced from pooled infected *Anopheles arabiensis* ovaries using the short paired-end BGI-DNBSeq technology. The reads were strin-

gently assessed for quality and cleaned up before de novo genome assembly was conducted. Genome analysis was done to identify genome completeness, followed by the prediction and annotation of key genes found in this symbiont. The assembled draft genome of *Microsporidia* sp. MB was found to have a total size of 5.9 Mb spanning 2,335 scaffolds, with an estimated 81% gene completeness compared to 600 core genes found in 13 microsporidian reference genomes. The symbiont has an average GC content of 31.12%. It contains 2,247 predicted putative genes based on the GeneMark-ES intronless eukaryotic gene prediction tool—a comparative assessment of the common orthologs between *Microsporidia* sp. MB and other microsporidia in the Enterocytozoonida group showed that the genome of *Microsporidia* sp. MB is phylogenetically closely related to *V. corneae*, separate from the mosquito-microsporidians found in other clades. This chapter provides insights into the genetic makeup of *Microsporidia* sp. MB and describes key features of its genome. These findings set the pace for future studies of host-symbiont interactions, transcriptomics, and strain variation analysis to further understand the exact blocking mechanism of this unique microorganism.

3.2 Keywords

Symbiont; Microsporidia; Vector Biology; Biocontrol; Genomics; Next-Generation Sequencing; De novo Assembly

3.3 Introduction

Microsporidia is one of the most expansive taxa, with approximately 1,600 species identified so far [Murareanu et al., 2021, Bojko et al., 2022]. A significant proportion of these species infect invertebrates, with over half found in insects [Becnel and Andreadis, 1999, Fast and Keeling, 2005, Shang et al., 2015]. They have diverse effects on their insect hosts, often dependent on host and microsporidia species, and can significantly impact agriculture, insect population regulation, and the economy at large [Pan et al., 2018]. Although the development of microsporidia in

insects follows a similar life cycle as that of mammalian hosts, there are some notable differences in the fine structure of the parasite. Microsporidia transmission in insects can occur in several pathways, including direct contact with infective spores, ingestion, and maternal transmission [Becnel and Andreadis, 1999, S.C. et al., 1982, Goertz and Hoch, 2008]. As intracellular parasites, the development of microsporidia in the host is often localized to the host cell's cytoplasm, with specialized relationships observed between microsporidia and its host at the cellular level [Becnel and Andreadis, 1999]. Pathogenic microsporidia such as the silkworm-infecting *Nosema bombycis* and the honeybee pathogen *Vairimorpha apis* have been reported to have a detrimental effect on the populations of these agriculturally important insects [Becnel and Andreadis, 1999, Bjørnson and Oi, 2014]. In contrast, other microsporidia have been explored as biocontrol tools against agricultural pests. For instance, the pathogenic *Paranosema locustae* effectively controls grasshopper populations while *Nosema pyrausta* similarly controls the European corn borer [Becnel and Andreadis, 1999, Bjørnson and Oi, 2014]. Several microsporidia species have been found to infect mosquitoes, exhibiting high levels of host specificity and varying effects from pathogenic to symbiotic [Andreadis, 2007]. Species belonging to the genera *Nosema*, *Amblyospora*, *Hazardia*, *Encephalitozoon*, *Enterocytozpora*, and *Microsporidium* have been identified in different mosquito species [Andreadis, 2007, Trzebny et al., 2023]. Due to the extensive impact of microsporidia in mosquitoes, these parasites have great potential for use in vector populations and disease transmission [Becnel et al., 2005]. Pathogenic microsporidia of mosquitoes include *Amblyospora connecticus* isolated from the saltmarsh mosquito [Andreadis, 1988], *Nosema algerae* infecting *An. stephensi*, an invasive malaria vector recently found in Africa [Undeen and Alger, 1975], and *Parathelohania anophelis*, which affects several *Anophelines* including *An. quadrimaculatus* [Avery and Undeen, 1990]. Microsporidia such as *Edhazardia aedis* cause mortality and reduced reproductive fitness in *Aedes* mosquitoes and dictate the blood-feeding success of the arboviruses vector [Koella and Agnew, 1997]. *Vavraia culicis* and *Microsporidia* sp. MB isolated in *Anopheles*, have been found to impact plasmodium development in *An. gambiae* and *An. arabiensis* respectively [Bargielowski and Koella, 2009, Herren et al., 2020].

3.3.1 Microsporidia genome variations in structure are dictated by their mode of transmission

Microsporidia mold themselves to adapt effectively to their host environment. For this reason, these microorganisms exhibit significant genomic variations from species to species depending on the complex relationship with the host species, microsporidia biology of infection, and transmission routes. Modes of microsporidia transmission, especially vertical transmission, play an essential role in shaping the genome size, complexity, and proportion of protein-encoding sequences [Haag et al., 2019]. Microsporidia with vertical transmission undergoes recurrent bottlenecks, leading to a decrease in population size and reduced purifying selection, potentially leading to genomes with more repetitive elements and non-coding regions. On the other hand, microsporidia with multiple transmission modes (vertical and horizontal) often exhibit larger genomes and fewer gene features [Haag et al., 2019].

3.3.2 Current advances in Next-generation sequencing (NGS) technologies and their application in microsporidia genomics studies

NGS technologies offer rapid, high-throughput, and cost-effective genomic analysis [Lee et al., 2013, Rizzo and Buck, 2012]. These technologies have broad applications across various research fields, such as whole genome sequencing, environmental DNA analysis, and gene expression profiling [Lee et al., 2013, Shokralla et al., 2012, Luo and Chen, 2019]. BGI-DNBSeq is a cost-effective sequencing platform known for its high throughput of short-read sequences, typically producing 150 bp paired-end reads. While its affordability makes it suitable for large-scale projects, the short reads can pose challenges for genome assembly and resolving complex regions [Guo et al., 2016]. In contrast, Illumina's NovaSeq 6000 offers higher accuracy and can generate short reads up to 300 bp, making it a preferred option for whole-genome sequencing, RNA-seq, and metagenomics, and it is often regarded as the gold standard for short-read sequencing due to its balance of cost,

speed, and precision (Hu, 2024, Logsdon et al., 2020). For long-read sequencing, technologies like Oxford Nanopore and SMRT PacBio excel at resolving complex genomes and structural variants. Oxford Nanopore is capable of producing reads over 100 kb in real-time and is portable, which is advantageous for fieldwork, although its raw read accuracy is generally lower, with recent improvements in base-calling algorithms addressing this issue [Jain et al., 2016]. SMRT PacBio, on the other hand, delivers highly accurate long reads, known as HiFi reads, exceeding 15 kb. This makes it ideal for de novo genome assembly and resolving repetitive sequences, though it tends to be more expensive and has lower throughput compared to short-read platforms [Wenger et al., 2019].

One of the main challenges of current NGS technologies is the vast amount of data generated, requiring exhaustive computational resources for efficient analysis [Lee et al., 2013, Tang et al., 2015]. The overwhelmingly rapid evolution of NGS platforms, each using different complex chemistries and applications, can also be challenging to scientists constantly trying to play catch-up [Rizzo and Buck, 2012]. NGS technologies have revolutionized the study of the complex biological systems exhibited by microsporidia, providing useful insights into their diverse genetic makeup and host interaction mechanisms [Slamovits et al., 2004, Peyretailade et al., 2011]. This has also enabled research into identifying genes involved in immune evasion and metabolic pathways that contribute to the virulence and transmission of microsporidia [Pombert et al., 2012, Corradi, 2015]. Comparative genomic studies of different microsporidia species reveal conserved genome features unique in each species essential for host adaptation and pathogenesis, and studies on genetic variations that may influence host specificity and pathogenicity [Keeling et al., 2005].

DNBSeq is an NGS technology developed by the BGI genomics company. This technology uses the combinatorial probe-anchor synthesis (cPAS) and DNA nanoballs, which are small, circular DNA structures created by rolling circle replication. The process involves converting DNA fragments into DNA nanoballs which are then immobilized onto a flow cell for sequencing [Rajesh and Jaya, 2017]. This technology is particularly unique as it enables high-throughput sequencing with high accuracy and low error rates. Additionally, DNBSeq offers scalability and is cost-

effective, making it a competitive option in the field of NGS. Choosing between short-reads and long-reads sequencing platforms is usually a trade-off between quality and quantity. While the former produces good-quality reads with high and low error rates, assemblies from these platforms are often riddled with unresolved gaps. On the other hand, long-read sequencing platforms, such as Oxford Nanopore and PacBio, might generate reads as long as 200Kbs, but with poor coverage, the assembly is prone to inaccuracies. Combining the two techniques would be ideal and has been applied in hybrid assemblies, significantly improving assembly quality and quantity [Minei et al., 2018]. Short paired-end sequencing techniques enable detecting a wide range of genomic variations. Its high resolution and accuracy make it a valuable tool for research and clinical applications, significantly advancing our understanding of genome structure and function.

WGS workflows involve several steps, from sample preparation to the final sequence data analysis. These steps include extraction of high-molecular-weight genomic DNA, DNA fragmentation, library construction with adaptors ligated to the fragments, and sequencing these fragments from both ends. This approach generates high-quality overlapping reads that can efficiently be aligned to a reference genome or assembled de novo.

3.3.3 Overview of gene prediction tools

Gene prediction tools are essential for analyzing genomic sequences and identifying gene structures. The tools can be broadly categorized into ab initio and evidence-based methods, with some utilizing comparative genomics or expression data to refine predictions. There are multiple gene prediction tools freely available online, each of which is optimized for specific applications. The accuracy of these tools is species-specific and may vary across different genomes. MetaEuk is a reference-based toolkit used for high-throughput prediction and annotation of genes in eukaryotic metagenomic sequences [Karin et al., Levy Karin et al., 2020]. This tool is handy in annotating complex gene and genome structures of underrepresented unicellular eukaryotes in metagenomic datasets, which often have lower sequencing coverage and poorly described eukaryote-specific experimental and computational procedures. The software conducts fast searches with six-frame-translated frag-

ments covering all potential exons. It then optimally combines these matches to construct multi-exon proteins. MetaEuk has notable high sensitivity, detecting genes with even low sequence similarity to the reference database. However, it has some disadvantages, including its heavy reliance on reference databases, high computational demands, and novelty of predicted genes. Augustus is frequently used in a supervised manner and requires prior knowledge for training before prediction. However, it can also incorporate RNA-Seq read information to improve its predictions, which makes it a powerful tool for genome annotation [Stanke and Morgenstern, 2005]. Augustus prediction is improved when used together with GeneMark-ET in the BRAKER1 pipeline, making unsupervised RNA-Seq-based genome annotation possible. Nevertheless, it is important to note that the effectiveness of Augustus may be compromised without proper training sets or the use of RNA-Seq data [Stanke and Morgenstern, 2005]. Glimmer is an ab initio prediction tool used primarily in prokaryotic genome annotations. It is reported to have a high sensitivity and specificity. However, this tool is error-prone, producing a higher number of false predictions compared to other tools. GeneMarkS is a self-training gene prediction tool used in both prokaryotic and eukaryotic genomes [Besemer et al., 2001]. It is known for its high prediction accuracy in gene start prediction [Besemer et al., 2001]. GeneMarkS-2 is an updated version of the program that uses heuristic models for training initialization and adaptive model structure modification, further enhancing its performance [Besemer et al., 2001]. GeneMark-ET integrates RNA-Seq data for eukaryotic genome annotation, and GeneMark-EP uses genomic footprints of homologous proteins. The main advantage of the GeneMarkS family over other tools is their self-training capability, which enables them to be used on newly sequenced genomes without additional preparation steps [Besemer et al., 2001].

3.4 Methodology

Unless otherwise stated, the default parameters of all tools were used in this study. The snakemake workflow designed in this study is available in [Appendix 4](#).

3.4.1 Sample collection and DNA extraction

Microsporidia sp. MB is efficiently transmitted from an infected mother to its progeny [Herren et al., 2020]. Gravid female *A. arabiensis* mosquitoes were collected in Ahero (34.9190°W, -0.1661°N), Western Kenya, between 2021-2023 and used to establish single-family lines in the laboratory. Once these female mosquitoes laid eggs, DNA was extracted from the mothers and screened for *Microsporidia* sp. MB to confirm infection using MB18S primers as previously described [Herren et al., 2020]. Offspring from *Microsporidia* sp. MB-infected mothers were retained as potential sequencing candidates. Ovaries from these progenies were dissected, and each ovary was placed in an Eppendorf tube containing absolute alcohol (That is, one mosquito = two tubes, each containing an ovary). A single ovary per offspring was extracted and screened for *Microsporidia* sp. MB using standard PCR. All ovaries with confirmed infection were pooled, and high molecular weight genomic DNA was extracted similarly. Quantitative PCR assays were set up to determine the infection validity of the pooled ovaries using the same MB18S primers [Herren et al., 2020]. Highly infected samples were selected for genome sequencing after evaluation of DNA quality (more than 5ng/microliter) using Qubit Fluorometric Quantitation (ThermoFisher Scientific, Waltham, USA).

3.4.2 DNA sequencing and de novo assembly

Samples were sent to BGI Genomics in China (<https://www.bgi.com/global>) for sequencing. Paired short-insert libraries were prepared using KAPA HiFi Hot-Start Library Amp Kit and sequenced with DNBSeg™ technology (2 × 150 bp reads), generating 326,181,200 raw paired-end reads. Adapter trimming was done using Trimmomatic. Raw reads were filtered using SOAPnuke v2.1.8 [Chen et al., 2018] using quality filtering parameters: `"-n 0.001 -l 10 -q 0.5 -adaMR 0.25 -polyX 50 -minReadLen 100"` where `"-n 0.001"` means N (ambiguous nucleotides) rate threshold of 0.001, `"-l 10"` represents a low-quality threshold of 10, `"-q 0.5"` filters reads with low-quality rate threshold of 0.5, `"-adaMR 0.25"` trims the adapter if the sequencing read matches 25.0% or more of the adapter

sequence (maximum two base mismatches are allowed), `"-polyX 50"` discards a read with poly-X number not smaller than the 50 where X is any one base, and `"-minReadLen 100"` removes all reads shorter than 100bp. Quality control checks of the final raw reads were done using FastQC v0.11.9 [Andrews, 2010] and MultiQC v1.12 [Ewels et al., 2016] to confirm no adapters or contaminants were present in the dataset before analysis. As intracellular parasites, isolating pure microsporidia gDNA from host gDNA is often challenging. Therefore, host reads were removed by mapping the dataset to *A. arabiensis* (GenBank: GCA_000349185.1) and *A. gambiae* s.s. (GenBank: GCA_000005575.2) reference genomes using the Burrows-Wheeler Aligner (BWA-MEM) v0.7.17 [Li and Durbin, 2009] [*Anopheles* reference genomes retrieved from RefSeq, last accessed January 2023]. Unmapped reads were filtered from the output BAM files using samtools v1.3.1 [Danecek et al., 2021] and retained for subsequent steps. Bacterial decontamination of unmapped reads was done using Kraken2 v2.0.8 [Wood et al., 2019] against the minikraken_8GB_20200312 database. De novo assembly of the cleaned-up reads using Unicycler v0.4.9 [Wick et al., 2017].

3.4.3 Quality control and visualization of the assembled genome

An NCBI MEGABLAST search [Johnson et al., 2008] of the assembled contigs against the non-redundant (NR) database was done to confirm similarity to other microsporidian sequences (E-value cut-off = 1×10^{-5}). Contigs not similar to microsporidia-like sequences were discarded, and raw reads were realigned to the confirmed microsporidia contigs using BWA-MEM to produce consensus contigs. This improved coverage and cleared out contamination during de novo assembly. Sequences non-homologous to microsporidia-like sequences were excluded to focus on those most relevant sequences in the genome. This approach was validated to ensure that critical microsporidian sequences were retained, and extensive quality control measures were implemented to minimize the risk of omitting essential data. As a result, confidence was maintained that the integrity and relevance of the target sequences were preserved throughout the analysis. The genome sequences have been deposited at DDBJ/EMBL/GenBank (accession <https://www.ncbi.nlm.nih.gov/submit/seq/submitseq.cgi>).

[gov/nuccore/JAVKTW000000000.1](https://www.ncbi.nlm.nih.gov/nuccore/JAVKTW000000000.1); GenBank ID: GCA_032191455.1). SprayN-Pray v3 [Garber et al., 2022] was used for the identification and taxonomic profiling of the genome contigs against a custom database that contained multiple microsporidia, bacterial and invertebrate genomes. Qualitative analysis of the GC content of each assembled contig and average read coverage was done using the Quality Assessment Tools (QUAST) v5.2.0 [Gurevich et al., 2013]. Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.4.3 [Simão et al., 2015] was used to assess genome completeness. BUSCO analysis was done against OrthoDB v10 [Kriventseva et al., 2019] microsporidia_odb10 database ($n = 600$ core microsporidia genes) [Aurrecochea et al., 2011] [last accessed in December 2023]. The assembled genome was visualized on BlobToolKit v3.5.2 [Laetsch et al., 2020] as a snailplot highlighting the general genome statistics and a pie-chart of BUSCO analysis. Gene and protein prediction was done using Genemark-ES v2.5p [Besemer et al., 2001] in intronless eukaryotic mode.

3.4.4 *Microsporidia* sp. MB repeats and transposable elements annotation

The presence of nucleotide repeats was done on RepeatMasker v4.1.2 [Nishimura, 2000] and RepeatModeler v2.0.4 [Flynn et al., 2020] against the [RepBaseRepeatMaskerEdition-20181026](#) database.

3.4.5 Comparative phylogenomic analysis of *Microsporidia* sp. MB and the *Enterocytozoonida* group

OrthoFinder v2.5.4 [Emms and Kelly, 2015, 2019], Buscophylo v1.3 [Sahbou et al., 2022] and OrthoDB v10 [Kriventseva et al., 2019] were used to perform a BUSCO-based phylogenomic study of *Microsporidia* sp. MB and clade IV microsporidia. Genome data used for this analysis included: *V. corneae* (GCF_000231115.1), *Hepatospora eriocheir* GB1 (GCA_002087885.1), *Enterocytozoon bieneusi* H348 (GCF_000209485.1), *Enterocytozoon hepatopenaei* (renamed *Ecytonucleospora hepatopenaei*) (GCA_026692725.1), *Enterospora canceri* (GCA_002087915.1) and *Nosema granulosis* (GCA_015832245.1) [last accessed in December 2023]. BUSCO

predictions were performed on each genome using BUSCO v5.4.3 [Manni et al., 2021]. Alignments of the identified common orthogroups across the species were done using MUSCLE [Edgar et al., 2004] and trimmed using trimAl [Capella-Gutiérrez et al., 2009]. The maximum likelihood tree was inferred using IQ-TREE v1.6.12 [Nguyen et al., 2016] with the best-fit model selection from ModelFinder [Kalyaanamoorthy et al., 2017] using the following default parameters: `"-bb 1000 -alrt 1000 -nt AUTO -ntmax"`. *Nosema granulosis* was used as an outgroup. Tree visualization was done using Geneious Prime v2023.2 [Olsen, 2014].

3.5 Results and Discussion

3.5.1 Quality control analysis during and after sequencing

3.5.1.1 Quality control of raw reads

Being obligate intracellular parasites, isolating the microorganism from host cells is challenging. Rigorous decontamination steps are necessary to sequence and assemble microsporidia isolated from host tissues. A total of 323,951,584 paired-end reads were sequenced and passed the preliminary QC check. Each raw read was 150bp long and contained no adaptors after filtration using SOAP-nuke. FastQC v0.11.9 [Andrews, 2010] and MultiQC v1.12 [Ewels et al., 2016] further confirmed the quality of these raw reads pre- and post-decontamination and host-removal steps as highlighted (Figure 3.1a.)

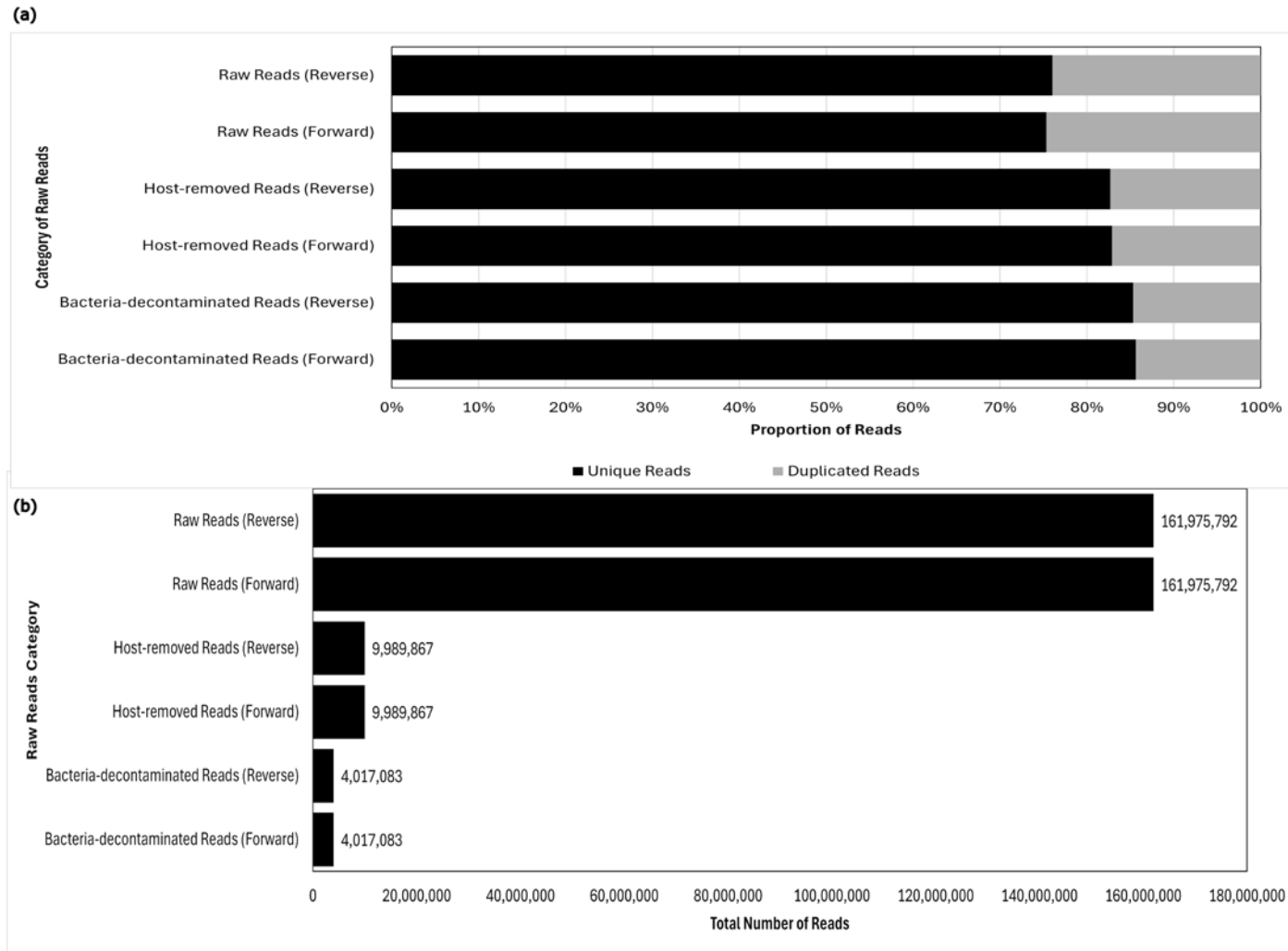


Figure 3.1: (a) The proportion of unique reads in each dataset based on FastQC analysis. (b) Number of total unique reads retained post-decontamination and host removal.

The number of reads reduced significantly after decontamination and host removal (Figure 3.1b). Classification of raw reads into their respective superkingdoms was done using Kraken2, revealing a higher proportion of unclassified reads (Figure 3.2, Table 3.1) followed by bacterial and eukaryotic reads. Archeal and viral reads were represented in a small proportion of the total reads (0.0026% and 0.0011%, respectively). All bacterial, archaeal, and viral reads were removed from the dataset before assembly.

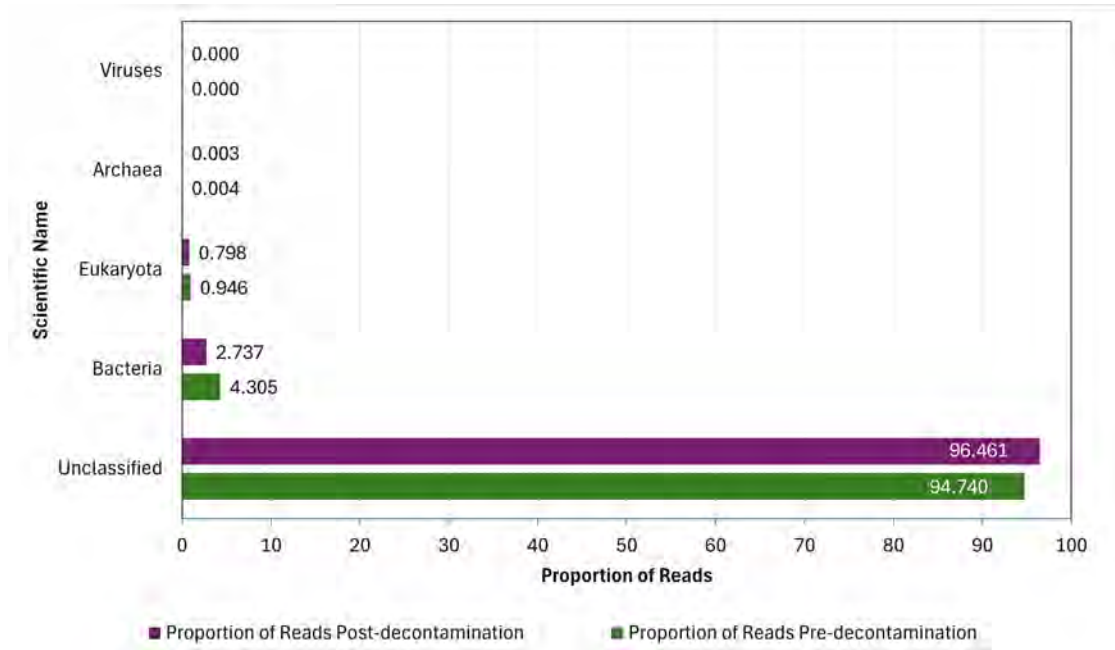


Figure 3.2: Superkingdom taxonomic classification of raw reads before and after decontamination using Kraken2. Only unclassified and eukaryotic reads were used for de novo assembly

3.5.2 *Microsporidia* sp. MB has a typical small genome

Microsporidia has the most petite eukaryotic genome sizes, which are comparatively similar to prokaryotic genomes, ranging from 2.3Mb (*Encephalitozoon intestinalis*) to 51.35 Mb (*Edhazardia aedis* USNM 41457) [Corradi et al., 2010, Desjardins et al., 2015]. A total of 3,874,899 clean raw reads were de novo assembled, resulting in 2407 contigs of lengths ranging from 26,338 to 200 bp, spanning a

Table 3.1: Distribution of classified raw reads before and after decontamination using Kraken2

Category	Number of Reads	Number of Reads
	Pre-decontamination (Relative Abundance)	Post-decontamination (Relative Abundance)
Unclassified	153,462,777 (94.74%)	3,874,899 (96.4605%)
Bacteria	6,973,106 (4.305%)	109,965 (2.7374%)
Eukaryota	1,532,450 (0.9461%)	32,070 (0.7983%)
Archaea	6,782 (0.0042%)	106 (0.0026%)
Viruses	677 (0.0004%)	43 (0.0011%)

total genome size of 5,908,979 bp. The genome statistical analysis *Microsporidia* sp. MB using QUAST v5.2.0 [Gurevich et al., 2013] showed that the genome contained 31.12%, within the range of most microsporidia species [Corradi and Selman, 2013]. A comparative analysis of the variable genome of *Microsporidia* sp. MB and *V. corneae* showed that the former had a larger genome spanning 2,407 contigs (2,335 scaffolds) with a slightly higher gene content of 2,247 compared to 2,239 *V. corneae* genes. Additionally, the symbiont contained a higher proportion of single-copy core genes (81%) than *V. corneae* as highlighted in Table 3.2.

3.5.2.1 Visualization of the *Microsporidia* sp. MB genome architecture using BlobToolKit

Visualization of the assembled genome general statistics as a snailplot using BlobToolKit v3.5.2 [Laetsch et al., 2020] highlighted the distribution of contigs, GC content, and gaps in the assembly. BUSCO scores were illustrated using a pie chart (Figure 3.3).

Table 3.2: Comparison of *Microsporidia* sp. MB with its closest relative *V. corneae* *ATCC 50505*

Metric	<i>Microsporidia</i> sp. MB	<i>V. corneae</i>
Assembly size (bp)	5,908,979	3,213,516
Number of contigs	2407	314
Contigs N ₅₀ (kb)	5	50.9
GC content (%)	31.12	36.5
The proportion of repeats (%)	0.57	Not available
Number of predicted genes	2,247	2,239
Gene density (genes/kb)	0.363	Not available
Mean CDS length (bp)	1,108	Not available
BUSCO (n=600)		
No. (%) of complete genes	486 (81.0)	450 (75)
No. (%) of complete and single-copy genes	485 (80.8)	443 (73.8)
Number (%) of complete and duplicated genes	1 (0.2)	7 (1.2)
Number (%) of fragmented genes	12 (2.0)	22 (3.7)
Number (%) of missing genes	102 (17.0)	128 (21.3)
GenBank accession number	JAVKTW000000000	NZ_- AEYK000000000
SRA accession number	SRR25938329	Not available
Sequencing technology	DNBSeq	454

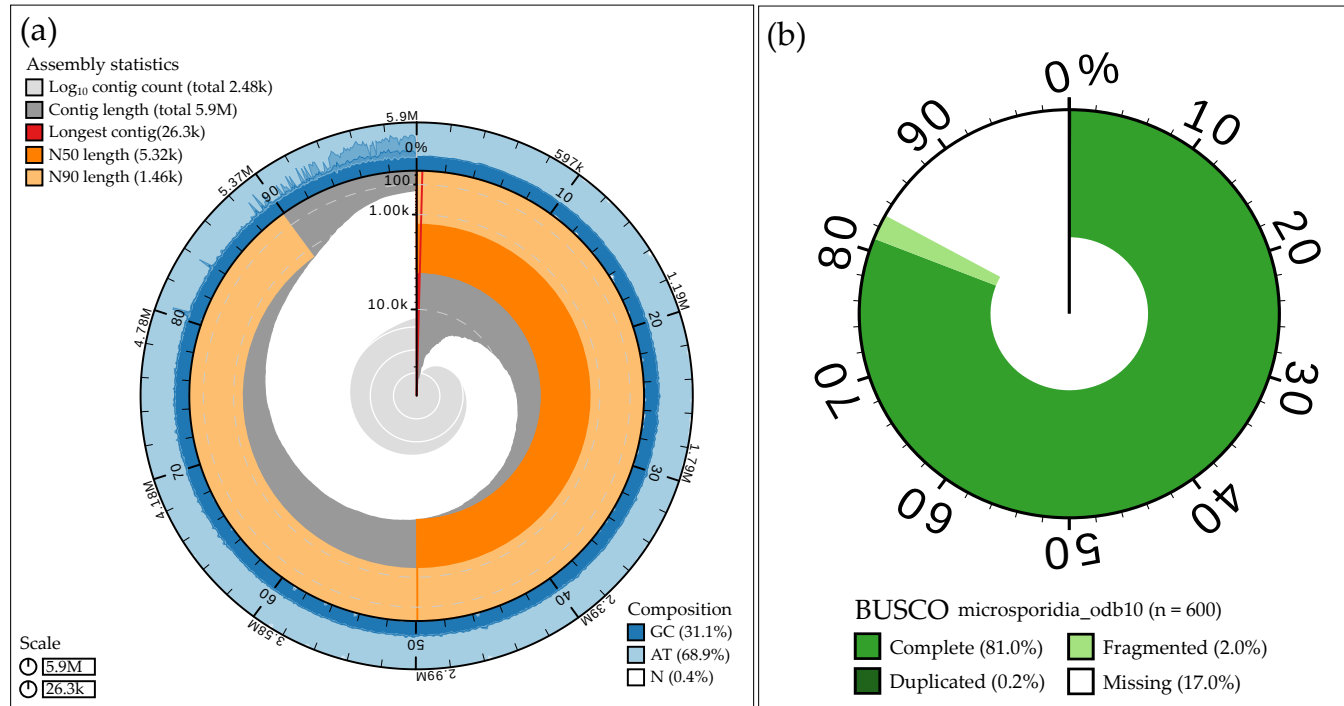


Figure 3.3: *Microsporidia* sp. MB genome structure. **(a)** The snail plot created using BlobToolKit highlights the genome metrics. A basic legend of the assembly statistics is provided in the top left corner. The outer blue rings highlight the GC content over the assembly, corresponding to each of the contigs in the grey ring. The overall GC content of the assembly is summarized at the bottom left corner. The light grey rings show the total length of the assembly. The red line highlights the longest scaffold, while the dark and light orange rings show the length of the contigs at which 50% and 90% of the total assembly are represented, respectively. **(b)** A pie-chart representation of the BUSCO gene completeness analysis demonstrates a total of 81% single-copy core microsporidia genes present in *Microsporidia* sp. MB.

3.5.2.2 Taxonomic classification of the assembled contigs reveal close homology to *V. corneae* and *N. ceranae* microsporidia species.

Taxonomic comparisons based on DNA provide complementary insights alongside protein-based phylogenetic analyses. While protein-based trees are generally more sensitive for assessing species relatedness due to the evolutionary conservation of protein sequences, DNA-based comparisons remain valuable. DNA data, particularly from non-coding regions or rapidly evolving sequences, can capture genetic variations not reflected in protein sequences. Together, DNA and protein-based analyses offer a broader view of genetic diversity and evolutionary history, providing a more comprehensive understanding of Microsporidia species relatedness. The phylogeny of microsporidia is constantly changing as studies reveal more ultrastructural and molecular information on these ever-evolving organisms [Park and Poulin, 2021]. *Microsporidia* sp.MB has been classified in the Clade IV set of microsporidia using the SSU rRNA region [Herren et al., 2020, Park and Poulin, 2021]. Taxonomic distribution of contigs using NCBI BLAST showed that the assembled contigs have a noticeably higher number of *V. corneae* BLAST Top-Hits (Figure 3.4) . All non-microsporidian contigs were removed from the assembly.

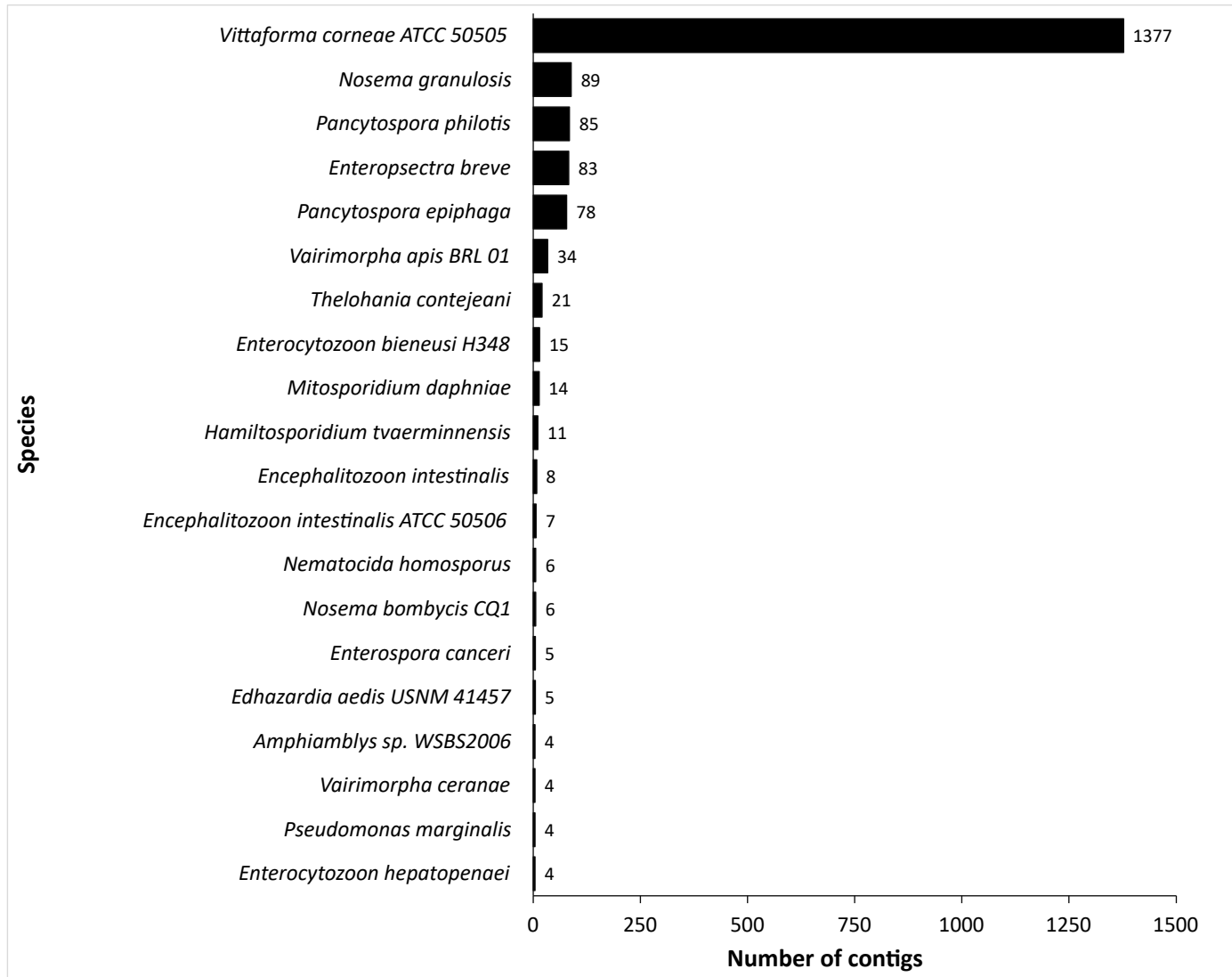


Figure 3.4: Species distribution of top BLAST hits

Taxonomic classification of the contigs using SprayNPray v3 [[Garber et al., 2022](#)] further confirmed their identity as microsporidia (Figure [3.5](#)).

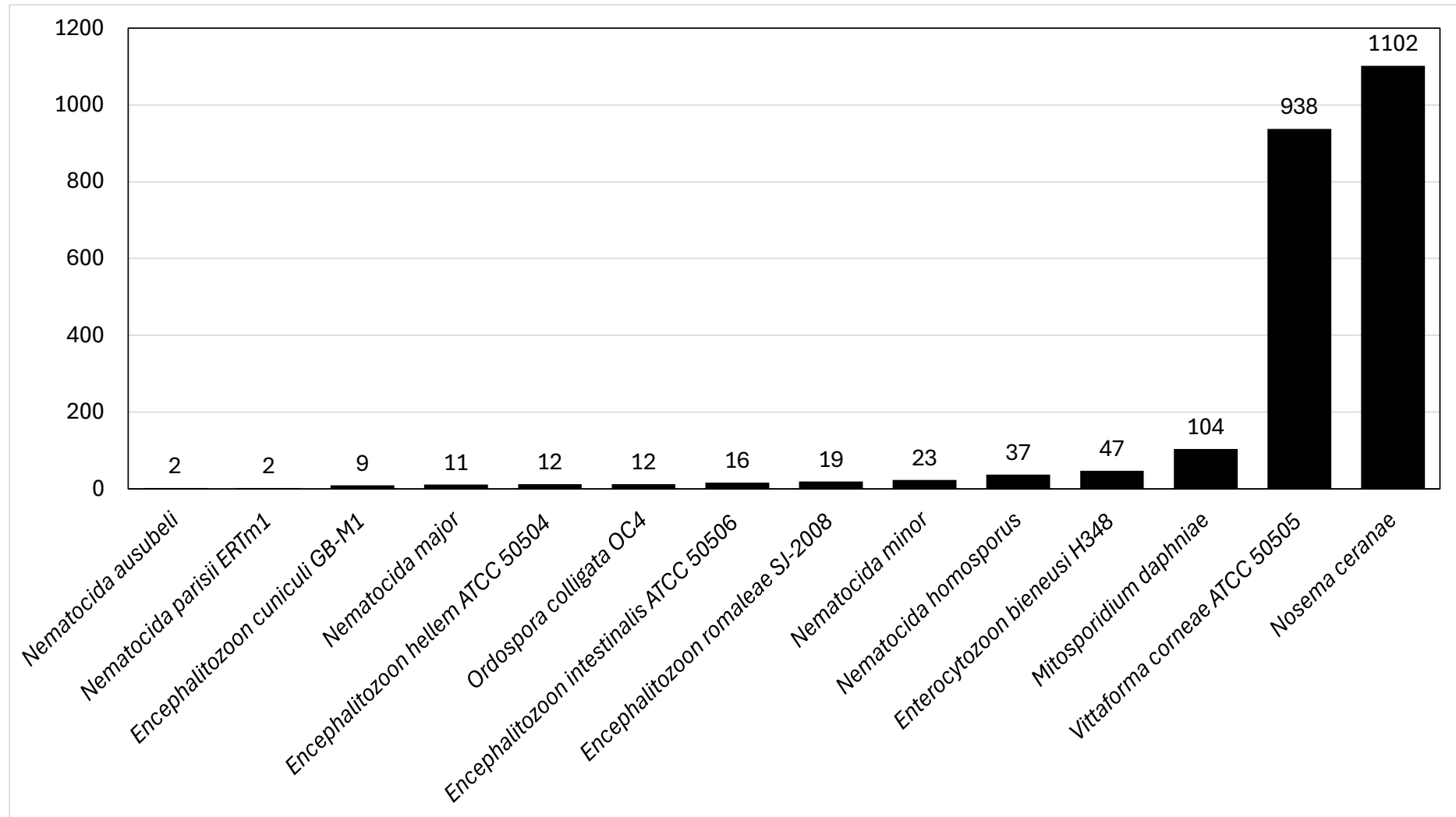


Figure 3.5: A homology search using SprayNPray v3 [Garber et al., 2022] identified all cleaned contigs were microsporidia. The most abundant species homologous to the *Microsporidia* sp. MB contigs were *N. ceranae* and *V. corneae*.

3.5.2.3 The small *Microsporidia* sp. MB genome could be attributed to the distribution of minimal repeats and transposable elements

Larger genome sizes are often attributed to large intergenic regions, high distribution of repeat regions, and low gene densities as seen in *Brachiola algerae* and *Edhazardia aedis* [Williams et al., 2008]. *Microsporidia* sp. MB contained minimal repeats (0.57%) as assessed using RepeatMasker v4.1.2 [Nishimura, 2000] and RepeatModeler v2.0.4 [Flynn et al., 2020] against RepBaseRepeatMaskerEdition-20181026 database (Table 3.3, Figure 3.6). Among the most frequently annotated repeat elements, the genome lacked short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), long tandem repeat elements (LTR), and DNA transposons.

Table 3.3: Distribution of different repeat types identified in the *Microsporidia* sp.

MB assembly		
Repeat Type	Total Number of Elements	Percentage of Sequence covered (%)
(SINEs)	0	0
LINEs	0	0
LTR elements	0	0
DNA transposons	0	0
Rolling circles	0	0
Unclassified	0	0
Small RNA	821	0.01
Satellites	0	0
Simple repeats	22,870	0.39
Low complexity	10,645	0.18
Non-repetitive	5,874,643	99.42

Several low complexity and simple repeats also called short tandem repeats (STRs), were found in *Microsporidia* sp. MB in low proportions (Figure 3.6). STRs have been linked to the stability and structural organization of microsporidia genomes through RNA repair and recombination, helping to maintain the genome integrity as seen in *Edhazardia aedis* [Williams et al., 2008, Gill et al., 2008].

They also play a role in regulating gene expression, and their characterization using transcriptomics analysis can be used to understand their host-symbiont interactions [Parisot et al., 2014]. Due to these repetitive elements' expansion and compaction and potential use as molecular markers, understanding genetic diversity subject to selective pressures in different hosts and environments could be possible [Heinz et al., 2012, Corradi and Slamovits, 2011].

A deeper analysis of these transposable elements in *Microsporidia* sp. MB could inform their roles in the symbiont as eukaryotic STRs may have multiple vital functions, including pathogenic consequences, translation, and transcription regulation [Wright and Todd, 2023].

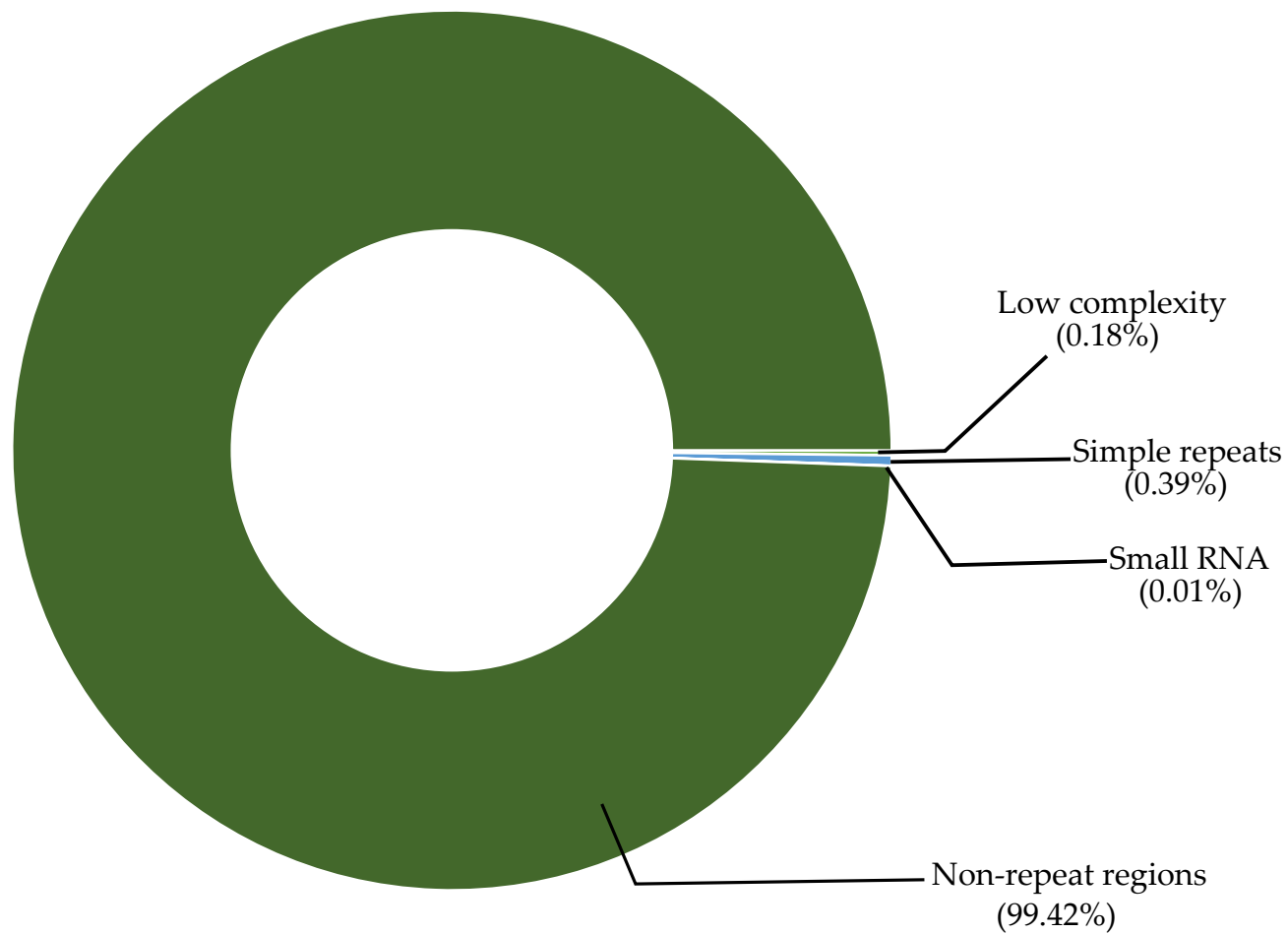


Figure 3.6: Distribution of different repeat types in *Microsporidia* sp. MB. The presence of short tandem repeats in small proportions was identified in this dataset.

3.5.2.4 The gene-dense *Microsporidia* sp. MB genome

A total of 2,247 genes were predicted using Genemark-ES v2.5p [Besemer et al., 2001]. Each gene was confirmed to be homologous to microsporidia genes through a similarity search on NCBI BLASTp (E-value cut-off value set at 0.0001). The organization of these putative genes corresponding to the BUSCO single-copy complete core genes was determined and schematically represented using the RStudio gggenes package. Figure 3.7 highlights the organization of genes on a subset of contigs of lengths greater than 10Kb, ranging from 26.338 - 10.051 Kb ($n = 48$). These genes were color-coded based on their general function based on BUSCO and GeneMark-ES annotation. Genes involved in metabolic and replication functions were highlighted in these top contigs.

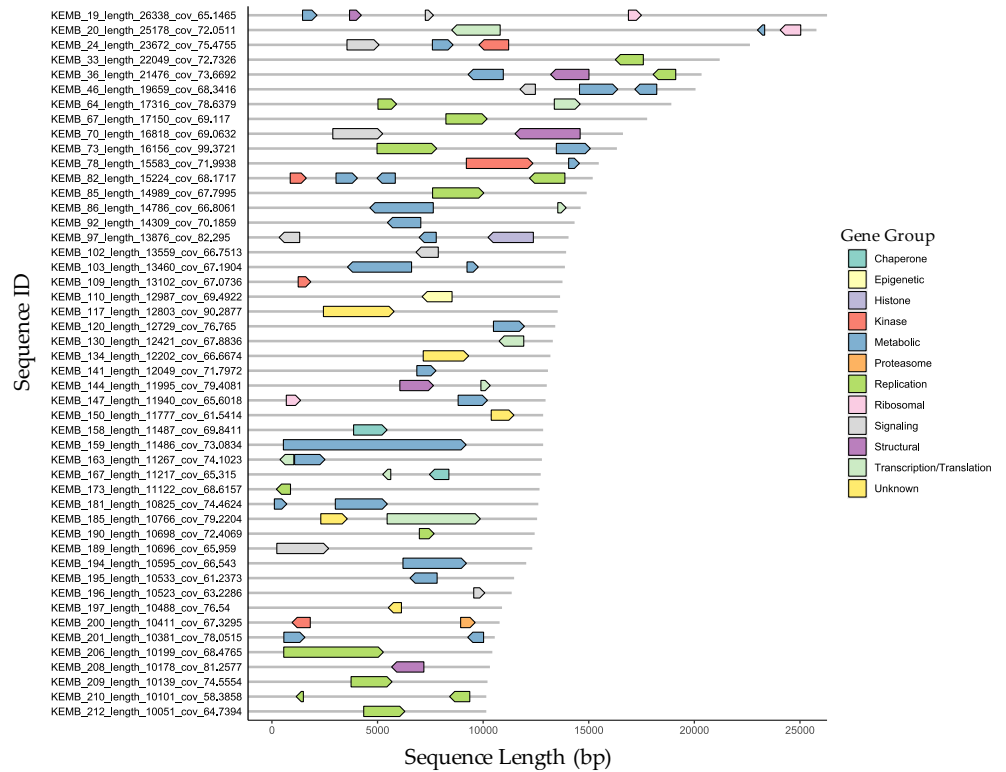


Figure 3.7: A representation of the genome structure of *Microsporidia* sp. MB, displaying contigs larger than 10Kb, highlighting gene density and organization per sequence. The identified genes correspond to the BUSCO identification of GeneMark-ES annotated proteins. A color-coded legend is included, categorizing the general function of each putative protein into 12 distinct categories. The plot was created in RStudio using the gggenes package (<https://cran.r-project.org/web/packages/gggenes>).

3.5.3 Comparative phylogenomic analysis of *Microsporidia* sp. MB and the Enterocytozoonida group places the symbiont closest to *V. corneae*

Phylogenomic analysis classified this symbiont as a close relative to *V. corneae* using BuscoPhylo v.1.3 [Sahbou et al., 2022]. Ninety-two common BUSCO orthogroups were identified across the genomes used in this comparative analysis (Figure 3.8). A concatenated tree using multiple sequence alignments from these orthogroups highlights the different clusters consistent with the 18S SSU rRNA taxonomical classification of clade IV microsporidia [Bukhari et al., 2022, Park and Poulin, 2021].

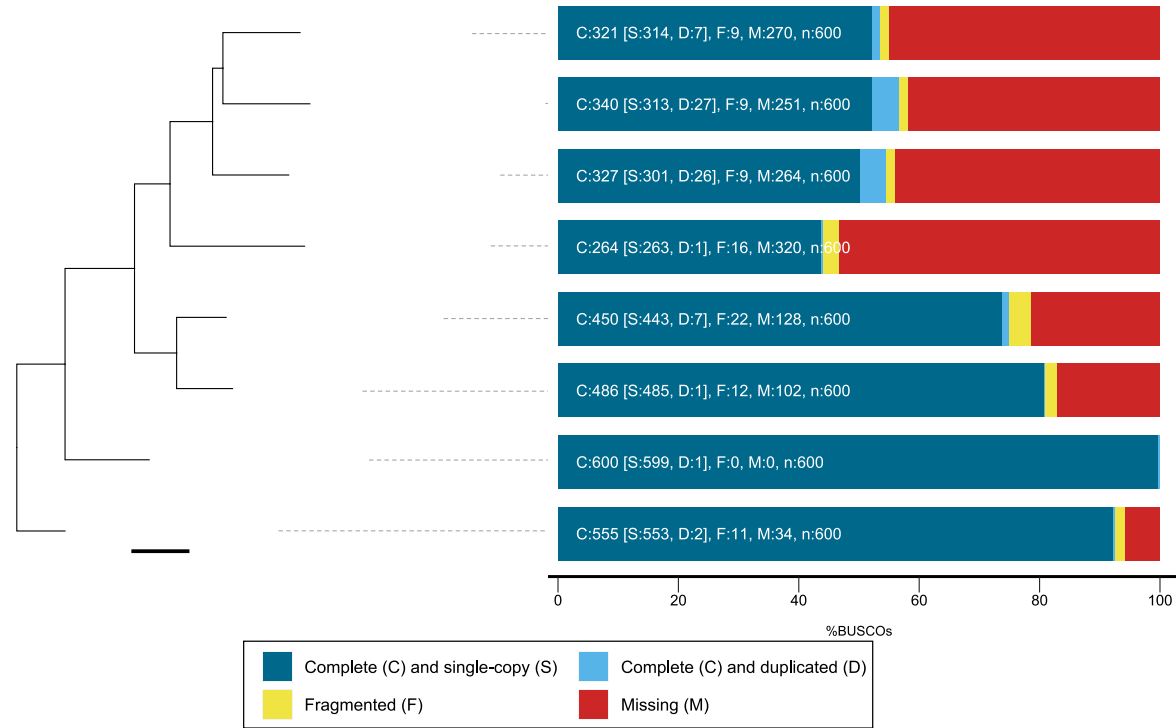


Figure 3.8: Phylogenomic analysis of common BUSCO orthogroups found in Enterocytozoonida (left) and their respective genome assembly completeness using BUSCO (right). *Microsporidia* sp. MB is in bold. *Nosema granulosis* was used as an outgroup. Common orthogroups were aligned using MUSCLE [Edgar et al., 2004] and trimAl [Capella-Gutiérrez et al., 2009]. The maximum likelihood tree (1,000 bootstraps) was constructed using IQ-TREE v1.6.12 [Nguyen et al., 2016] based on the 92 single-copy orthologous genes (with the best model selection (LG+F+I+G4) from ModelFinder [Kalyaanamoorthy et al., 2017]. Bootstrap percentage values are shown on each node, and the scale bar at the bottom of the tree indicates total changes per site.

3.6 Conclusion

Microsporidia consists of a group of highly-specialized eukaryotic organisms with the smallest genomes due to excessive gene loss over time owing to their over-reliance on the host for most of its resources to survive within the host cell. *Microsporidia* sp. MB has been isolated from *Anopheles* mosquitoes across Africa, showing promising potential as a malaria biocontrol tool. This chapter aimed to assemble the genome of this complex symbiont. Here, we highlight the steps involved in sample preparation, sequencing, and de novo assembly of the *Microsporidia* sp. MB genome. The genome is relatively small, at 5.9 Mbs, containing minimal repeats and 2,247 predicted genes. A total of 81% complete single-copy microsporidia core genes are present in the genome. The symbiont has a larger genome size than the other species in the Enterocytozoonida group, with the highest content of core BUSCO genes in the group. Establishing a reference genome for this symbiont is the first step in understanding its infection biology. It sets the stage for future host-symbiont-parasite interaction studies alongside other NGS and RNASeq experiments. The limitations of the BGI-DNBSeq platform, particularly its short 150 bp reads, were fully acknowledged in this study. While it was recognized that longer reads from third-generation sequencers would have been advantageous for generating more contiguous assemblies, challenges in obtaining high-quality tissue DNA for long-read sequencing, coupled with budget constraints, influenced the choice of this method. Despite the shorter reads, efforts were made to compensate by employing a robust assembly strategy aimed at maximizing genome coverage. Moreover, concerns regarding the identification of core genes were addressed by considering the relatively short gene lengths in microsporidia, including *Microsporidia* sp. MB. This characteristic allows for the recovery of complete genes, even from contigs shorter than the average gene length found in larger eukaryotes. The 486 core genes identified were verified as complete by comparison with conserved orthologous genes in microsporidia. Although 12 genes were identified as fragmented, this represents only a small fraction of the total, affirming the overall integrity of the assembly for the majority of core genes. With respect to using *V. corneae* as a reference for genome assembly, this was

initially considered. However, due to poor assembly results, this approach was abandoned in light of the significant diversity within microsporidia, even among closely related species. While *V. corneae* and *Microsporidia* sp. MB are within the same clade, their genomic differences are substantial. Microsporidian genomes are highly divergent, often displaying marked differences in gene content, structure, and repetitive elements, even between closely related species. Therefore, relying on *V. corneae* for scaffolding could have introduced errors due to these genomic disparities. The use of RagTag and mapping tools such as Bowtie2 will be considered for future assembly refinement.

Chapter 4

Microsporidia sp. MB Genome Annotation and Functional Analysis

4.1 Chapter Overview

The genome of *Microsporidia* sp. MB has recently been sequenced, providing a valuable resource for understanding the biology of this mosquito-infecting organism (See Chapter 3). This chapter uses various computational methodologies to examine the gene content and protein function of this symbiont. The analysis covers the distribution of protein function, physicochemical properties, conserved domains, and metabolic pathways. *Microsporidia* sp. MB has a compact genome of 5.9 Mbp, consisting of 2,247 genes, most of which are enzymes. It has retained all the glycolytic pathway-associated genes, unlike its counterparts in the Enterocytozoonida group. The genome also contains genes involved in the RNA interference pathway, suggesting a probable mechanism for suppressing host immune responses against the symbiont. The mode of sexual reproduction through meiosis is also studied here by annotating several meiotic-related genes in the genome, illustrating the haploid nature of the microsporidia. These findings provide insight into the general biology of *Microsporidia* sp. MB and highlight the dynamics of its genetic makeup. Further transcriptomic analyses of these genes could provide information on the relationship between the host and symbiont,

conferring the unique malaria-blocking effect.

4.2 Keywords

Genome Annotation; Metabolic Pathways; Symbiont; Biological Control; Vector; meiotic-related genes; glycolytic pathway; RNA interference

4.3 Introduction

4.3.1 Microsporidia genomes: models of extensive gene loss driven by their obligate intracellular nature

Microsporidia are highly specialized obligate intracellular microorganisms that have undergone extreme genomic and cellular reduction [Keeling et al., 2005, Haag et al., 2014, Slamovits et al., 2004]. These obligate intracellular parasites have lost a lot of genes involved in the synthesis of resources and have retained a large set of genes involved in the transportation of resources from the host cell [Keeling et al., 2010, Pombert et al., 2012, Nakjang et al., 2013, Wiredu Boakye et al., 2017].

An important characteristic of Microsporidia is the apparent absence of conventional mitochondrial functions such as haem biosynthesis and aerobic respiration [Burri et al., 2006]. This unique trait sets them apart from other eukaryotes and highlights the specialized nature of their intracellular lifestyle. Microsporidia contain small mitochondrial remnants known as mitosomes in place of mitochondria [Waller et al., 2009, Dolgikh et al., 2011, Burri et al., 2006]. They have a rudimentary protein import pathway, and the absence of mitochondria suggests that some microsporidia lack ways to synthesize ATP, relying heavily on host resources.

4.3.2 Multifaceted computational annotation reveals important protein functions

Using computational tools to characterize the structure and functions of unknown proteins has been widely described [Ang'ang'o et al., 2022, Islam et al., 2015, Desler et al., 2009]. These tools have been applied in studies on drug discovery

and vaccine development and have shown high prediction efficacy [Nyamai and Bishop, 2020, Brown, 2017, Boruah et al., 2013]. A recent example of a successful attempt to annotate a microsporidian genome is demonstrated by Svedberg et al. (2024), who developed a workflow combining sequence and structure-based approaches to annotate the *Vairimorpha necatrix* genome, significantly improving the accuracy of functional annotations [Svedberg et al., 2024].

4.3.3 Decay of components of the glycolytic pathway in microsporidia

Microsporidia have developed methods to obtain energy from the host cells and manipulate the metabolism of the host to acquire nutrients. The energy is obtained either directly from the host or through glycolysis during the spore stage [Wiredu Boakye et al., 2017]. The first genome sequenced from Enterocytozoonida, a major family of microsporidians, seemed to have lost the genes for glycolysis [Wiredu Boakye et al., 2017]. Based on a comprehensive analysis of gene content across the clade, the entire lineage of microsporidia, including *Enterocytozoon hepatopenaei* and *Enterospora canceri*, has lost most genes involved in the glycolytic pathway and lacks any obvious intrinsic means of generating energy. The reason for this loss of members of the glycolytic pathway in microsporidia is their highly specialized and reduced biology as obligate intracellular parasites [Keeling and Fast, 2002].

4.3.4 The RNA interference (RNAi) machinery

The RNAi mechanism involves four main stages and relies on three main proteins: the Dicer, Argonaute, and RNA-dependent RNA polymerase (Figure 4.1

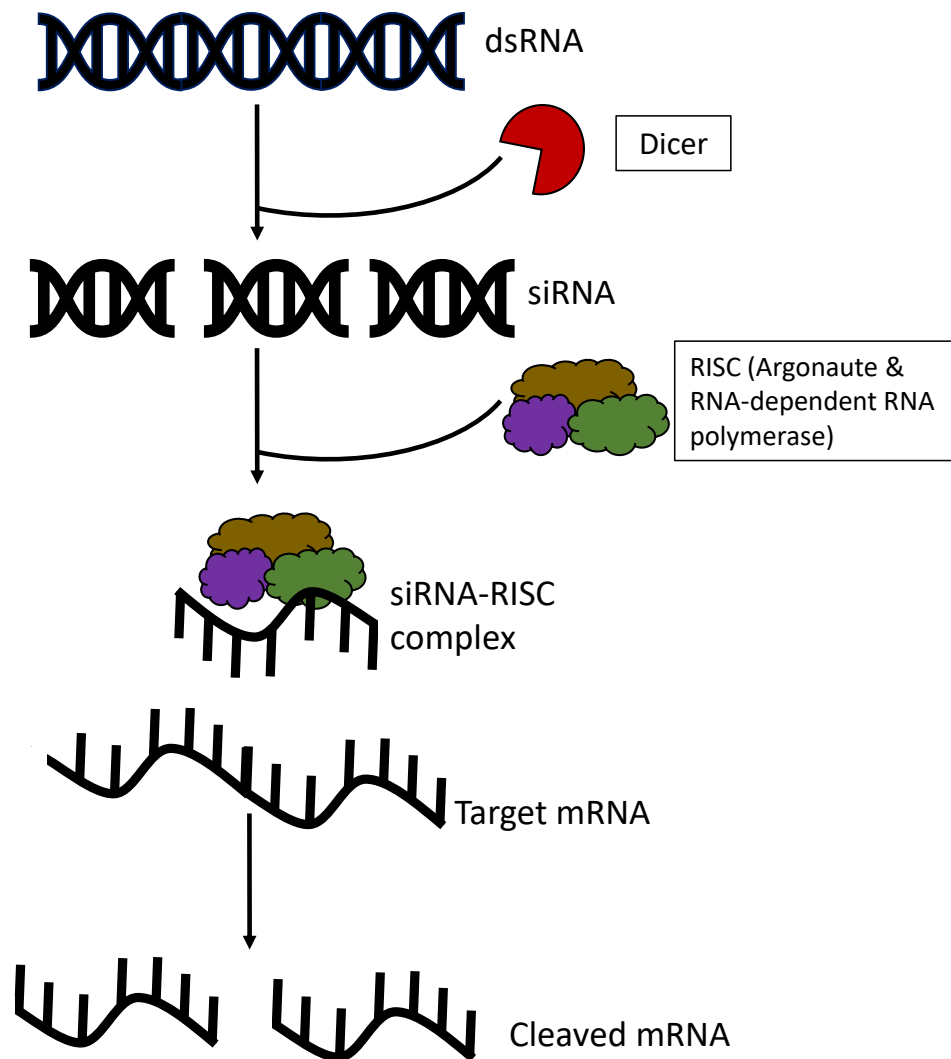


Figure 4.1: A schematic representation of the RNAi machinery highlighting the steps where the Dicer, Argonaute and RNA-dependent RNA polymerase have active roles in.

This machinery plays a crucial role in gene regulation and silencing and could be involved in parasite survival and pathogenesis [Paldi et al., 2010, Huang, 2018]. The RNAi process involves the production of small interfering RNAs (siRNAs) by an enzyme called Dicer, which cleaves RNA duplexes into short oligomers [Swanton et al., 2004]. The Argonaute then guides an endonuclease called the RNAi-induced silencing complex (RISC) to target and silence specific mRNA molecules [Swanton et al., 2004]. The RNAi pathway in microsporidia, including the gene Dicer and Argonaute orthologs, is selectively maintained or lost during lineage divergence, as observed in studies conducted by Huang [Huang, 2018] and Whitten *et al.* [Whitten et al., 2016].

4.3.5 Meiotic related genes

Meiosis in microsporidia is a sexual reproduction process that involves different stages during sporogony leading to spore production [Loubès, 1979]. Proof of meiotic processes occurring in some microsporidia is characterized by synaptonemal complexes in some species. Studies in the conservation of meiotic genes in eukaryotes can inform their mode of sexual reproduction and evolutionary relationships. The presence of meiotic genes in different organisms, ranging from protists to mammals, highlights their critical role in sexual reproduction [Malik et al., 2008, Tvedte et al., 2017, Ramesh et al., 2005]. The molecular evolution of meiosis genes has been studied in different organisms, including *Arabidopsis* and mammals and has been linked to epigenetic mechanisms [Wright et al., 2015, Yang et al., 2011, Yamaguchi et al., 2012]. The regulation of meiotic gene expression during meiosis in fungi such as fission yeast contributes to the description of the meiotic developmental pathway by identifying genes that are essential for meiosis and spore formation [Nag and Axelrod, 1998].

4.3.6 The Iron-Sulfur (Fe-S) cluster assembly machinery

Fe-S clusters are essential co-factors for numerous proteins involved in cellular processes such as metabolic reactions, gene expression regulation, and electron

transfer [Lill et al., 2006]. In eukaryotes, the Fe-S clusters are transported into apoproteins through a complex and highly conserved process involving multiple assembly machinery [Pandey et al., 2019, Lill et al., 2006]. Fe-S biogenesis pathways involve four main steps. The process starts with sulfur liberation by a cysteine desulfurase enzyme, forming a persulfide intermediate on a conserved cysteine residue. Iron, the other essential component, is donated by iron-binding proteins like frataxin and integrated into the cluster during assembly. The sulfur and iron components are assembled on scaffold proteins with the aid of electron donors, which are crucial for reducing sulfur to sulfide and forming the Fe-S cluster. Once the Fe-S cluster is complete, it is transferred to apoproteins via chaperone proteins. Chaperones ensure the proper assembly of the cluster onto the apoprotein, forming the mature holoprotein [Dellibovi-Ragheb et al., 2013]. The Fe-S cluster assembly machinery in eukaryotes is responsible for the biogenesis of these Fe-S co-factors, synthesizing Fe-S on scaffold proteins before transporting them to target apoproteins [Lill et al., 2006]. This machinery is vital in the maturation of mitochondrial, cytosolic, and nuclear Fe-S proteins [Pandey et al., 2019, Lill et al., 2006, Tsaousis et al., 2014].

4.4 Methodology

4.4.1 Overview of the annotation process

Details on the software and tools used for annotation are highlighted in the List of Tools ???. Unless specified otherwise, default parameters were applied to all tools.

4.4.2 Protein classification

Protein functional and structural classification was done using InterProScan [Blum et al., 2021]. This tool links several protein classification databases for optimal annotation, including Pfam, NCBIfam, and SUPERFAMILY, for protein family classification. CDD, PROSITE, and for conserved functional domain analysis.

4.4.3 Gene Ontology (GO) analysis

GO analysis [Ashburner et al., 2000] was conducted using BLAST2GO [Conesa et al., 2005] where BLAST results from each putative gene were mapped onto the predicted GO term for accuracy.

4.4.4 KEGG Metabolic pathway analysis

The distribution of key pathways and analysis of the genes involved in different pathways were done using KEGG [Kanehisa et al., 2016].

Visualization of the distribution of the different GO terms (biological processes (BP), molecular functions (MF), cellular components (CC), and KEGG pathways was done using RStudio 2023.09.1+494 [R Core Team, 2017].

4.4.5 Transmembrane localization and signal peptide predictions

Following the procedure highlighted in Section 2.4.2.2, TargetP v2.0 and SignalP v6.0 were used to identify signal peptides and cleavage sites, while DeepTMHMM v1.0.24 was used to predict transmembrane regions in the protein.

4.4.6 Physicochemical properties

ExPASy Protparam [Gasteiger et al., 2005a] was used to identify the physicochemical properties of the putative proteins as described in Section 2.4.2.3.

4.4.7 Conserved domains, motif analysis and homology modeling

Conserved domain analysis, motif analysis, and homology modeling were done as described in Section 2.4.3.3.

4.4.8 Annotation of LRR proteins

Leucine-rich repeat (LRR) proteins are characterized by the presence of repetitive sequences rich in leucine, forming a structural motif that is involved in various cellular functions, including ligand binding, pathogen recognition, and immune responses [Zhang et al., 2022]. In the context of Microsporidia, LRR proteins may be involved in interactions with the host immune system and could play a role in facilitating the invasion and survival of the parasite within host cells. While the specific functions of LRR proteins in Microsporidia remain underexplored, these motifs have been shown to play crucial roles in related parasitic organisms. For example, in *Phytophthora sojae*, they are highlighted to be crucial in development and virulence [Zhang et al., 2022]. Similarly, in *Eriocheir sinensis*, LRR proteins have been implicated in immune responses, highlighting their role in pathogen defense [Shi et al., 2024]. Based on the annotation of LRR proteins from InterProScan repeats analysis, genes identified to have these sections were selected from the dataset, and batch motif analysis was used to highlight LRR regions. Homology modeling and mapping motifs to the 3D structure were done as described in Section 2.4.3.3. Identification of signal peptides and transmembrane regions was done as in Section 2.4.2.2.

4.4.9 Characterization of the glycolytic pathway in *Microsporidia* sp. MB

A set of previously documented microsporidia proteins involved in the glycolytic pathway were identified using an extensive literature search [Tokarev et al., 2019, Luo et al., 2021, Wiredu Boakye et al., 2017, Vivarès and Méténier, 2008, Keeling and Corradi, 2011] and retrieved from the NCBI database [last accessed in December 2023]. *Encephalitozoon cuniculi* has been extensively studied and has retained all these genes and was thus used as a reference for identifying homologous genes in *Microsporidia* sp. MB.

The genes from *E. cuniculi* used in this study included: pyruvate kinase (ECU09_0640), enolase (ECU10_1690), fructose 1,6-bisphosphate aldolase (ECU01_0240),

phosphoglycerate mutase (ECU10_1060), hexokinase (ECU11_1540), pyruvate DHase E1, a sub (ODPA SCHPO) (ECU09_1040), pyruvate DHase E1, b sub (ODPB RICPR) (ECU04_1160), phosphofructokinase (ECU03_0680), phosphoglucose isomerase/glucose-6-phosphate isomerase (ECU05_0650), 3-phosphoglycerate kinase (ECU05_0320), glyceraldehyde-3-phosphate dehydrogenase (ECU07_0800) and triose phosphate isomerase (ECU11_0230).

A remote BLASTp search for homologous sequences in this protein dataset was done (E-value cutoff = 0.0001; bit-score greater or equal to 200), and the top 10 hits were retrieved for each protein. Reverse BLAST of the top 10 hits (see Section 2.2) was performed to confirm true homology. A custom local database was designed with the BLAST+ "makeblastdb" command [for Biotechnology Information, US], which included homologous proteins from the Enterocytozoonida group *V. corneae* ATCC 50505, *E. bienersi*, *E. hepatopenaei*, *E. canceri*, *H. eriocheir* and other microsporidia such as *N. granulosis*, *V. ceranae*, *A. locustae*, *O. bayeri*. Homology search of *Microsporidia* sp. MB genome against this custom database of glycolysis proteins was done. Multiple sequence alignment of the homologs was done using MUSCLE [Edgar et al., 2004]. ModelFinder [Kalyaanamoorthy et al., 2017] was used to identify the best-fit model for tree construction. Tree visualization was done in Geneious Prime® 2023.2.1 [Olsen, 2014].

4.4.10 Characterization of Meiotic-related genes (MRGs) orthologs in *Microsporidia* sp. MB

MRGs are integral to meiosis, the process critical for sexual reproduction and the generation of genetic diversity. These genes encode proteins that facilitate key meiotic events, including homologous chromosome pairing, recombination, and the accurate segregation of chromosomes into gametes, ensuring genomic stability across generations [Villeneuve and Hillers, 2001]. In the context of microsporidia, a group of obligate MRGs play a significant role in understanding their unique reproductive strategies. . Despite the streamlined nature of their meiotic processes, certain microsporidia species retain core meiotic genes, which are essential for their lifecycle and parasitism [Corradi and Selman, 2013]. Characterizing

MRGs in microsporidia is critical for elucidating their evolutionary adaptation and reproductive mechanisms in the context of their obligate parasitic lifestyle [Desjardins et al., 2015]. MRGs of microsporidia were retrieved from NCBI using references from published articles [Lee et al., 2014, Peat, 2011, Ndikumana et al., 2017]. A custom database using the assembled MB genome was created locally using the "makeblastdb" command-line argument. Retrieved MRGs were used as queries against this database using tBLASTn (E-value cutoff set at 1×10^{-20}). A reverse BLAST of all contigs from each hit was done by running a remote blast on NCBI. Reverse BLAST is a technique used to identify true homologous sequences by applying a more stringent search filter on the results of an initial BLAST analysis. In this process, homologous hits are re-evaluated using an in-house script that performs a secondary BLAST search. This script refines the results by focusing on hits where the query sequence appears among the top 3 matches, with criteria such as high coverage ($>50\%$), percent identity ($>40\%$), and e-value (<0.005). By applying these rigorous filters, the technique ensures that only true homologs are identified, reducing the chance of including sequences that are similar due to random chance rather than genuine homology. The putative genes predicted from GeneMarkS from each respective contig were also queried to confirm true homology. In this case, another custom BLAST database was created using the retrieved MRG protein sequences, and a local protein BLAST (BLASTp) search was conducted (E-value = 0.0001). The homology of the identified genes was further confirmed using remote BLAST against the nr database. InterProScan v2.1 [Blum et al., 2021] was used to identify functional domains. Multiple sequence alignment was done using MUSCLE (Sequences were ordered based on similarity). Alignments were colored by similarity – the darker the residue, the higher the similarity. Tree calculations using Fasttree and the Jukes-Cantor Genetic Distance Model were done with 1000 bootstrap replicates.

4.4.11 Annotation of the RNAi mechanism in *Microsporidia* sp. MB

The RNAi mechanism consists of 3 main proteins: Dicer, Argonaute, and RNA-dependent RNA polymerase [Whitten et al., 2016]. Homologs specific to each gene

were retrieved from MicrosporidiaDB v10 [Aurrecochea et al., 2011] [last accessed in December 2023]. The search and subsequent analysis of the *Microsporidia* sp. MB respective homologs to these three genes were done as described in Sections 4.4.9 and 4.4.10 above. These three genes have a set of unique conserved domains (PAZ Argonaute-like, DUF, and PIWI-like, Rnc, and Dicer domains), which were confirmed as described in Section 2.4.3.3.

4.4.12 Phylogenetic analysis of Iron-Sulfur cluster assembly proteins in *Microsporidia* sp. MB

Microsporidia genes involved in the Iron-Sulfur Cluster Assembly as reported in *T. hominis* were retrieved from NCBI [Williams et al., 2002]. To create a local database, a remote NCBI BLASTp search of these protein sequences was done to retrieve ten maximum hits (E-value less than or equal to 1e-20; query coverage greater than or equal to 90%). Putative genes from *Microsporidia* sp. MB was blasted against this custom database as described above. Multiple sequence alignments and tree calculations were conducted on the homologs using MUSCLE and FastTree, respectively.

4.5 Results and Discussion

4.5.1 Proteins involved in enzymatic processes within the cell form the most significant proportion of the genome

Of the 2,247 proteins predicted using BUSCO and GeneMark-ES tools, 2,238 were classified into their respective superfamilies using InterProScan [Blum et al., 2021]. (Figure 4.2). Over 30% ($n = 707$) of these proteins were associated with enzymes, and only 12 proteins (0.53%) were left unclassified. The other protein family groups identified included: cell-cycle ($n = 530$; 23.6%), regulatory/cellular ($n = 364$; 16.2%), binding ($n = 353$; 15.7%), transport ($n = 178$; .9%), and structural proteins ($n = 103$; 4.6%)(Figure 4.2) .

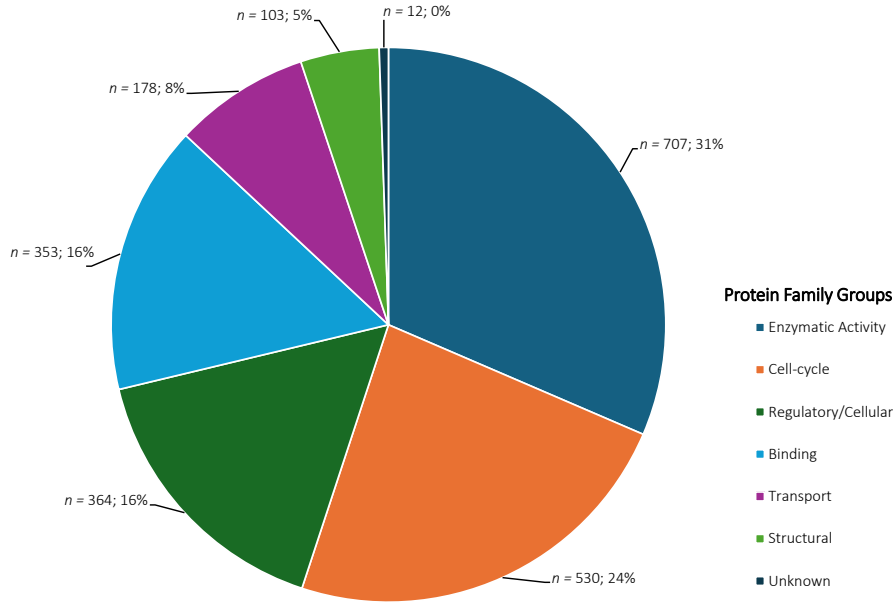


Figure 4.2: A pie-chart representation of the proportion of different protein family categories in *Microsporidia* sp. MB using InterProScan. Seven main protein superfamily groups were identified, with most of the proteins classified as enzymes.

4.5.1.1 Hydrolases and transferases are the most abundant enzyme classes in the *Microsporidia* sp. MB genome

Since most predicted proteins were classified as enzymes, further analysis of the different groups of enzymes was essential to determine the distribution of different enzyme classes represented in the dataset. Seven main enzyme classes were identified, with transferases appearing to be the most common category among the annotated enzymes (Figure 4.3). These were divided into: hydrolases ($n = 237$; 38.47%), transferases ($n = 228$; 37.01%), ligases ($n = 45$; 7.31%), isomerases ($n = 35$; 5.68%), oxidoreductases, ($n = 33$; 5.36%), translocases ($n = 31$; 5.03%) and lyases ($n = 7$; 1.14%).

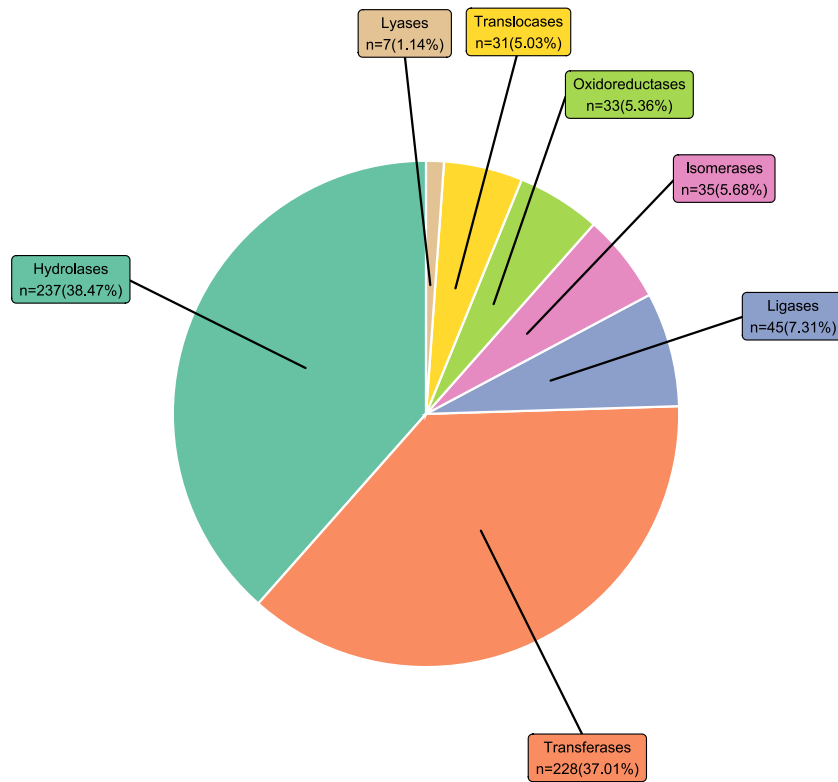


Figure 4.3: The distribution of primary enzyme groups using InterProScan, highlighting hydrolases and transferases as the most abundant enzyme classes in *Microporidia* sp. MB. These two enzymes function in energy metabolism and transport of nutrients respectively.

4.5.1.2 Distribution of different enzyme sub-classes

The enzyme classes identified were further grouped into sub-classes based on their functions (Figure 4.4).

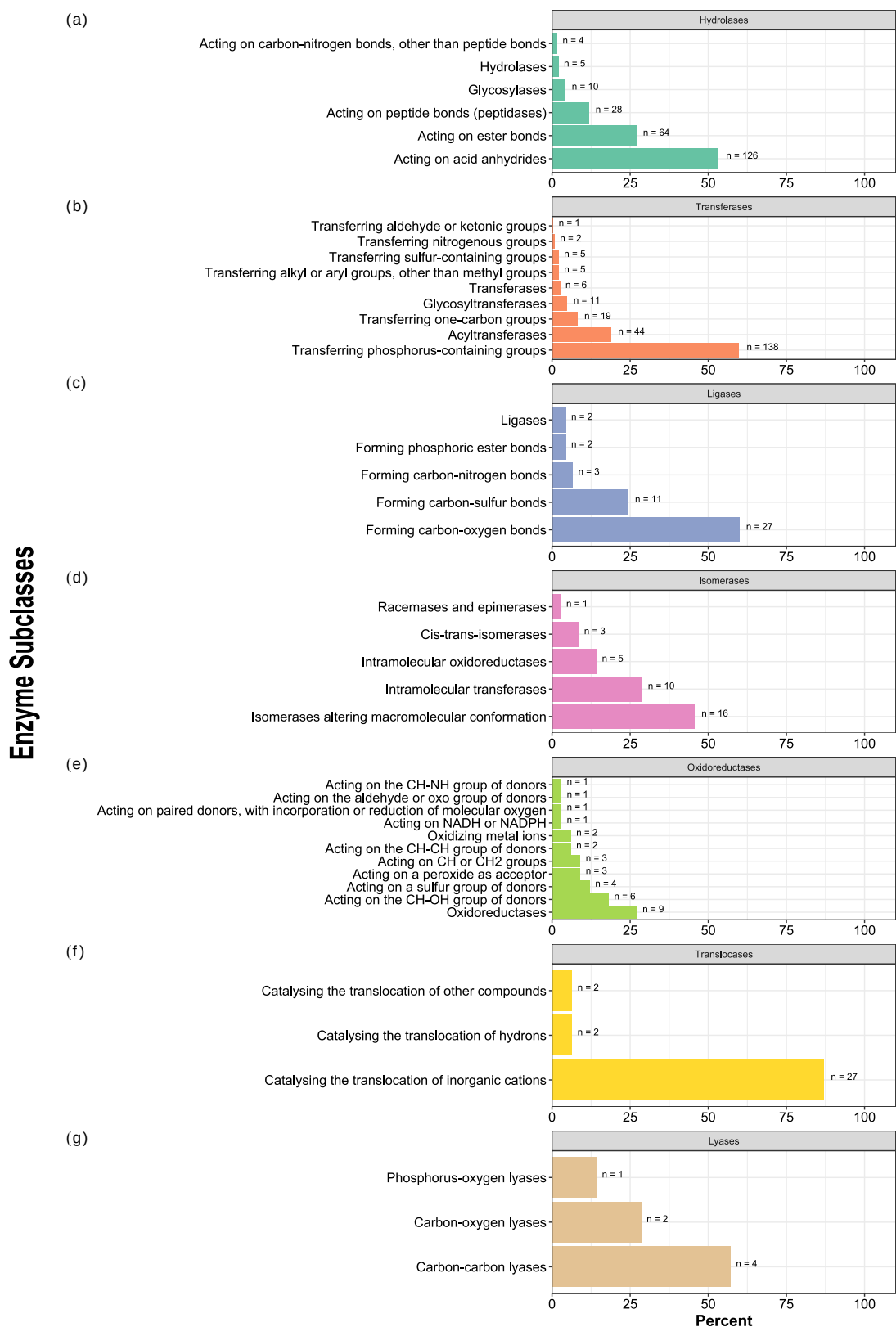


Figure 4.4: Distribution of different enzyme classes and sub-classes in *Microsporidia* sp. MB

Hydrolases are involved in the carbohydrate and energy metabolism of intracellular stages of the microsporidian *Nosema grylli* [Dolgikh, 2000]. The localization and functionality of microsporidian iron-sulfur cluster assembly proteins have been reported to be hydrolases [Goldberg et al., 2008]. A total of 38.47% of the classified enzymes in *Microsporidia* sp. MB were identified as hydrolases, with most of these acting on acid anhydrides (Figure 4.4a).

Transferases play a crucial role in microsporidia intracellular lifestyle, particularly in acquiring essential nutrients and energy from their host cells [Alexander et al., 2016, Dean et al., 2016]. These transporters also facilitate the import of nucleotides for nucleic acid biosynthesis. Nine main sub-classes of transferases were identified, with those involved in transferring phosphorus-containing molecules being the most common (Figure 4.4b).

The identification of several components of the mitochondrial iron-sulfur cluster assembly machinery in these organisms further supports the role of ligases in microsporidia. These components, which are involved in the biosynthesis of iron-sulfur proteins, are essential for various cellular processes, including DNA repair. The presence of these components suggests that microsporidia may use ligases to maintain their genome integrity and ensure their survival within the host cell. *Microsporidia* sp. MB contains 7.31% ligases of the predicted enzymes, which were further classified into five sub-classes depending on the type of bonds they create with different molecules (Figure 4.4c).

Oxidoreductases in microsporidia are not well-studied; however, their general physiological roles include involvement in hormone control, iron uptake, defense against pathogens, and toxin degradation through the redox system. A total of eleven oxidoreductase sub-classes were observed. Isomerases, oxidases, and thiol-sulfide reductases are important in redox reactions in the organism's survival in new host environments. Five isomerase sub-classes were identified, some directly linked to oxidoreductases and transferases. On the other hand, lyases are involved in various aspects of the microsporidian life cycle, including energy metabolism and nucleotide acquisition. These were the least-represented enzyme classes in the genome consisting of carbon-carbon, carbon-oxygen, and phosphorus-oxygen subgroups (Figure 4.4d).

4.5.2 *Microsporidia* sp. MB contains multiple Leucine-rich repeat (LRR) proteins

4.5.2.1 Distribution of domains and repeats in *Microsporidia* sp. MB

Protein domains, repeats, and conserved sites are essential features of proteins. Domains in proteins are distinct regions that have specific structural and functional properties, are typically composed of several secondary structure elements, such as alpha helices and beta sheets, and can fold independently. Repeats, conversely, are segments of protein sequences repeated multiple times within the same protein or across different proteins. Repeats can be indicative of functional motifs or structural elements within proteins. The predicted proteins contained a total of 659 domains and 19 repeats. A total of 1456 domains were identified in the putative proteins, with a majority being the reverse-transcriptase domains ($n = 179$) as shown in Figure 4.5. *Microsporidia* lack nucleotide biosynthetic pathways and rely on "stealing" nucleotides from host cells [Alexander et al., 2016, Dean et al., 2016].

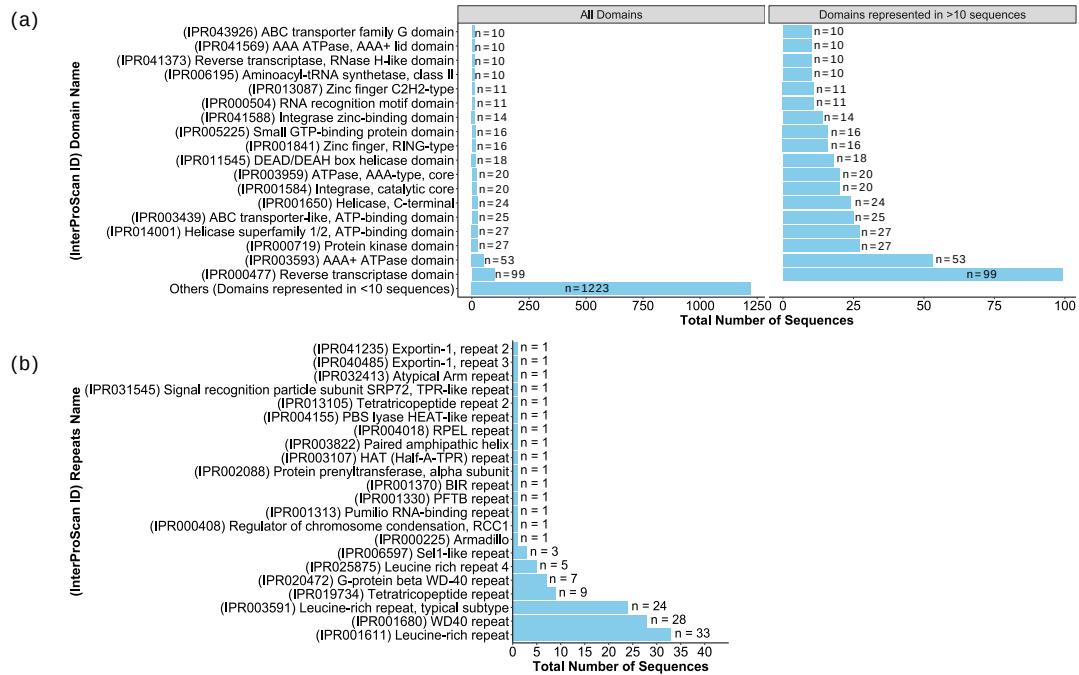


Figure 4.5: Sequence distribution of the different domains and repeats predicted using InterProScan. (a) A bar plot of the key domains isolated from the genome showing a large set of domains involved mainly in enzymatic activities. A zoomed image of the main domains excluding the outlier (where less than 10 sequences were represented in each domain) is shown on the right panel. (b) The distribution of repeat regions is depicted in this barplot. Leucine-rich repeat regions (LRR) and WD40 were the most abundant.

4.5.2.2 Characterization of LRR proteins in *Microsporidia* sp. MB

Most notable of the identified repeats was the presence of LRRs, with a total of 17 proteins containing these repeats (Figure 4.5b above). The LRRs found in microsporidia proteins play a crucial role in recognizing and interacting with host cells and may represent pathogenicity factors. These regions often involve mediating protein-protein interactions, allowing microsporidia to bind to and invade host cells [Senderskiy et al., 2014]. Microsporidian LRR-proteins secreted into the host have the potential to interact with host protein function. LRR-proteins in *S. lophii* were shown to have N-terminal signal peptides indicating potential secretion or targeting to the parasite cell surface for interactions with

the host [Campbell et al., 2013]. Similar sequences in *V. corneae* also underscore the significance of LRR-proteins in mediating protein-protein interaction [Campbell et al., 2013]. An investigation of the distribution of LRRs in the proteome using MEME Suite showed that 9 of the assembled contigs contained key LRR motifs (Figure 4.6). To identify whether the LRR-containing proteins could be part of the MB secretome, an essential feature in host-parasite interactions, further computational assessment of the presence of signal peptides and transmembrane localization prediction was done. Three of the identified contigs contained putative signal peptides, suggesting they could be part of the secretome, in direct contact with the host cell. Nine sequences with LRR regions were identified using InterProScan as highlighted in Section 4.5b. Motifs containing these repeats in the sequences were identified using MEME. One contig (NODE_405) contained two main LRR genes (gene_679 and gene_680) with confirmed signal peptides using both TargetP and SignalP prediction tools, while contig NODE_1665 contained one gene (gene_1719) (Figure 4.6). Gene_679 was identified to be homologous to the *Anopheles Plasmodium*-responsive LRR protein APL1B, which has been shown to play a role in *An. gambiae* immune response to invasion by *Plasmodium* [Williams et al., 2015] and could suggest a probable function of gene_679 in blocking malaria in its host.

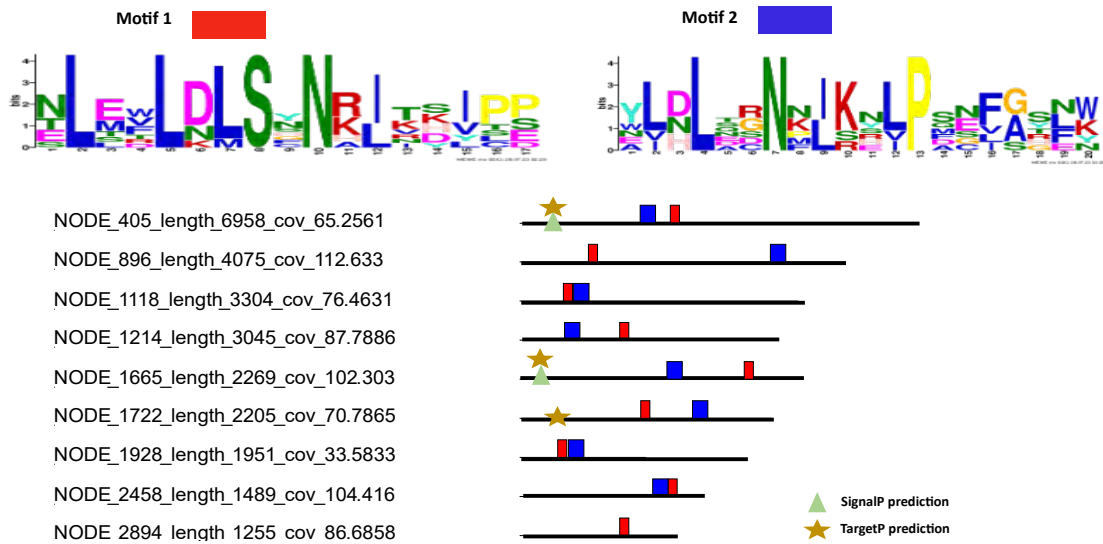


Figure 4.6: Identification and analysis of leucine-rich repeats (LRR) motifs across 9 contigs. LRR motifs 1 (red) and 2 (blue) were identified in the nine contigs using MEME Suite. Moreover, prediction of signal peptides using TargetP v2.0 and SignalP v6.0 (green and brown respectively) in 2 of these sequences (NODE_405 and NODE_1665).

The 3D model of the putative protein encoded by gene_679 was designed using template 3O6N_A APL1 (1.85Å) and was characterized with eight alpha helices and 11 beta sheets and loops (Figure 4.7). Quality scores of the modeled proteins included a QMEAN score of 0.5, Verify3D overall score of 77.98%, and Z-DOPE score of -3.83. Gene_679 also contained signal peptides and transmembrane segments (Figure 4.6). No good-quality template was identified from HHSearch to model gene_680 and gene_1719.

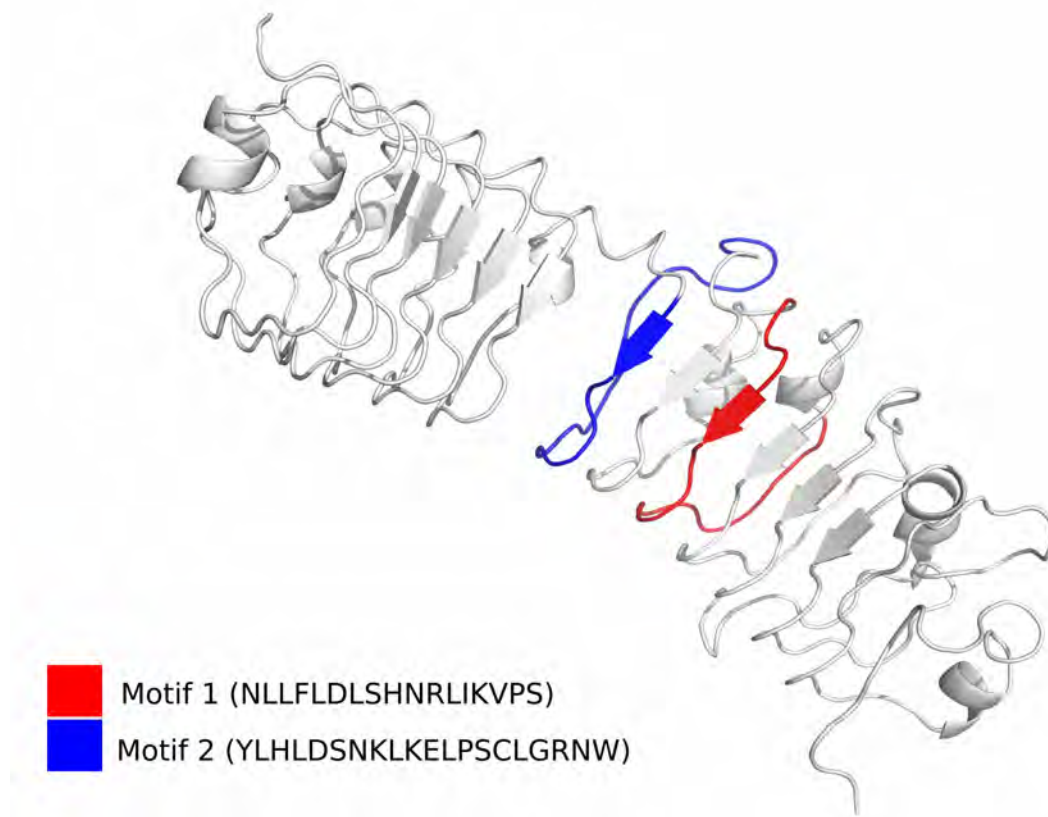


Figure 4.7: A 3D model of gene_679 (NODE_405) with motif mapping. The structure was modeled on PRIMO. The identified LRR motifs are numbered based on the MEME output and color represented is consistent with Figure 4.6 above.

4.5.3 *Microsporidia* sp. MB proteins are mainly involved in metabolic processes

4.5.3.1 GO analysis

GO analysis is a commonly used method to annotate genes and gene products and to determine the biological functions of large-scale genome or transcriptome data [Ashburner et al., 2000]. In this context, GO analysis showed that a majority of the putative proteins in *Microsporidia* sp. MB are involved in metabolic processes and transport, which can be attributed to the bulk of enzymes found in the predicted dataset. Most of these proteins were found in the nucleus and cytoplasm (Figure 4.8).

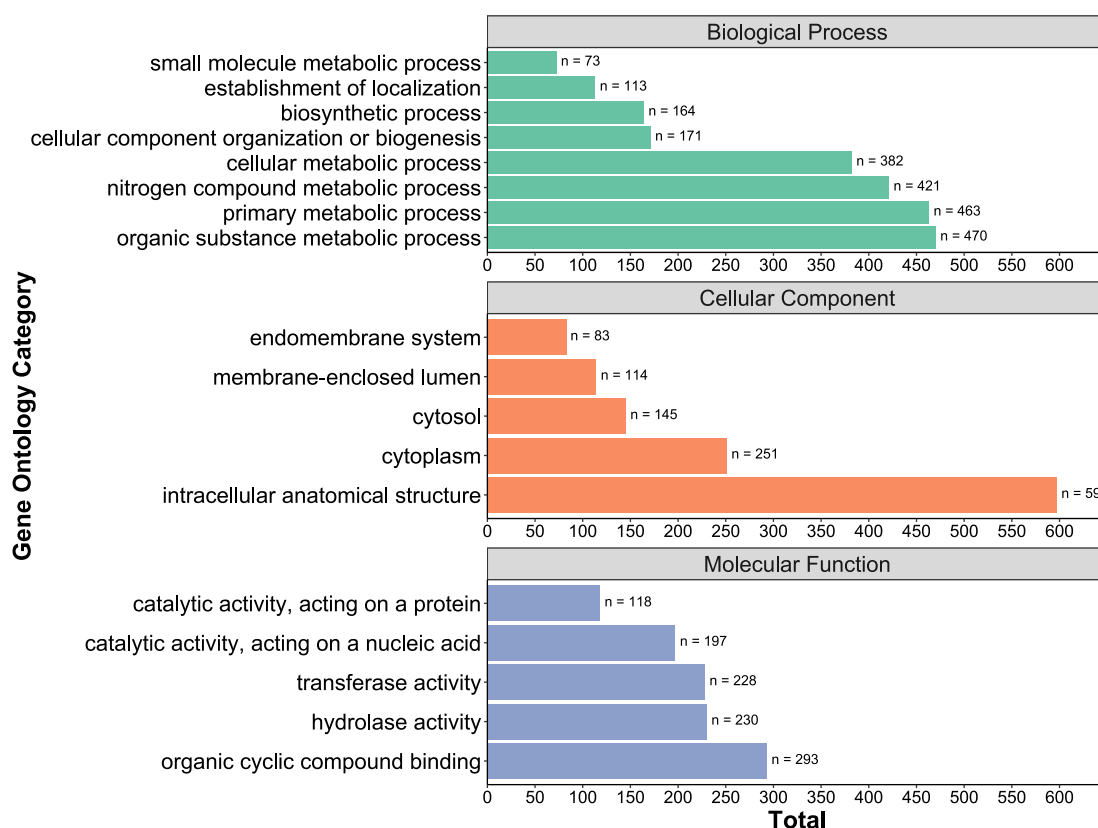


Figure 4.8: A highlight of the protein sequence distribution in each primary GO term (Biological Process, Cellular Component and Molecular Function) shows that a bulk of the putative proteins are involved in metabolic processes, localized within the intracellular anatomical structure (nucleus and cytoplasm) and involved in enzymatic activities.

4.5.3.2 KEGG pathway analysis

A total of 257 pathways were identified, with 741 sequences linked to these pathways (E-value = 0.0001). KEGG further categorized the proteins into five distinct groups. Most of the proteins were involved in metabolism ($n = 83$; 32.3%) and organismal systems ($n = 83$; 32.3%) (Figure 4.9). The other categories included environmental information processing ($n = 35$; 13.6%), cellular processes ($n = 30$; 11.7%), and genetic information processing ($n = 26$; 10.1%).

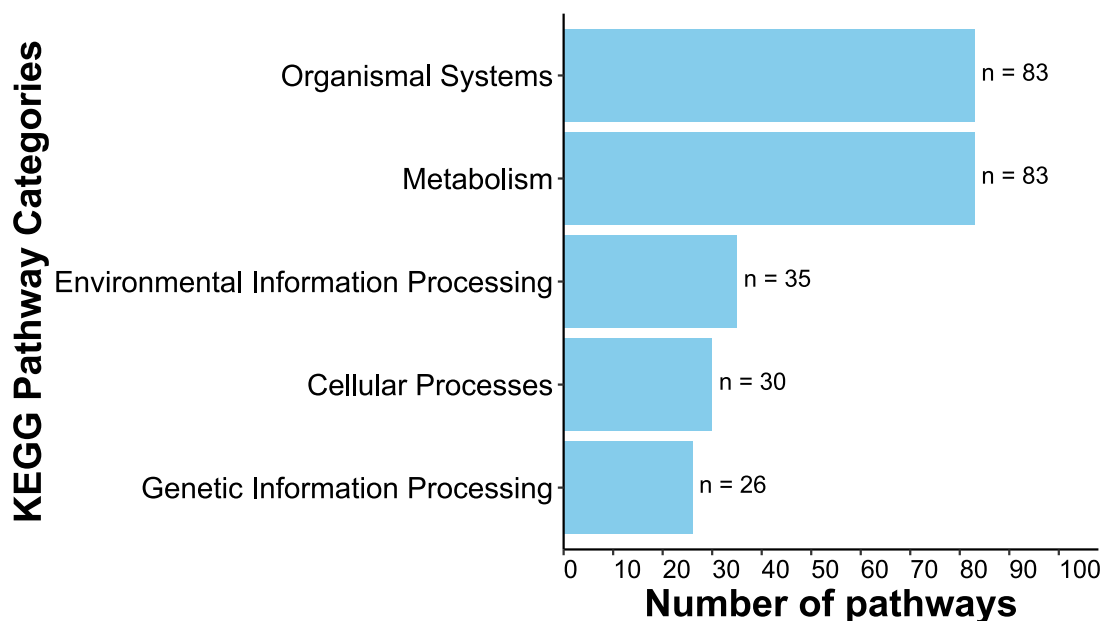


Figure 4.9: KEGG metabolic pathway analysis. A barplot representation of the main KEGG categories shows most of the pathways identified within the genome are involved in organismal systems and me-tabolism.

Most of the predicted proteins are involved in metabolism, genetic information processing and general cellular processes, including purine, thiamine, and nucleotide metabolism, biosynthesis of co-factors, ribosome biogenesis, protein processing, nucleotide excision repair, RNA degradation, RNA polymerase, cell cycle, meiosis, cellular senescence and nucleocytoplasmic transport (Figure 4.10).

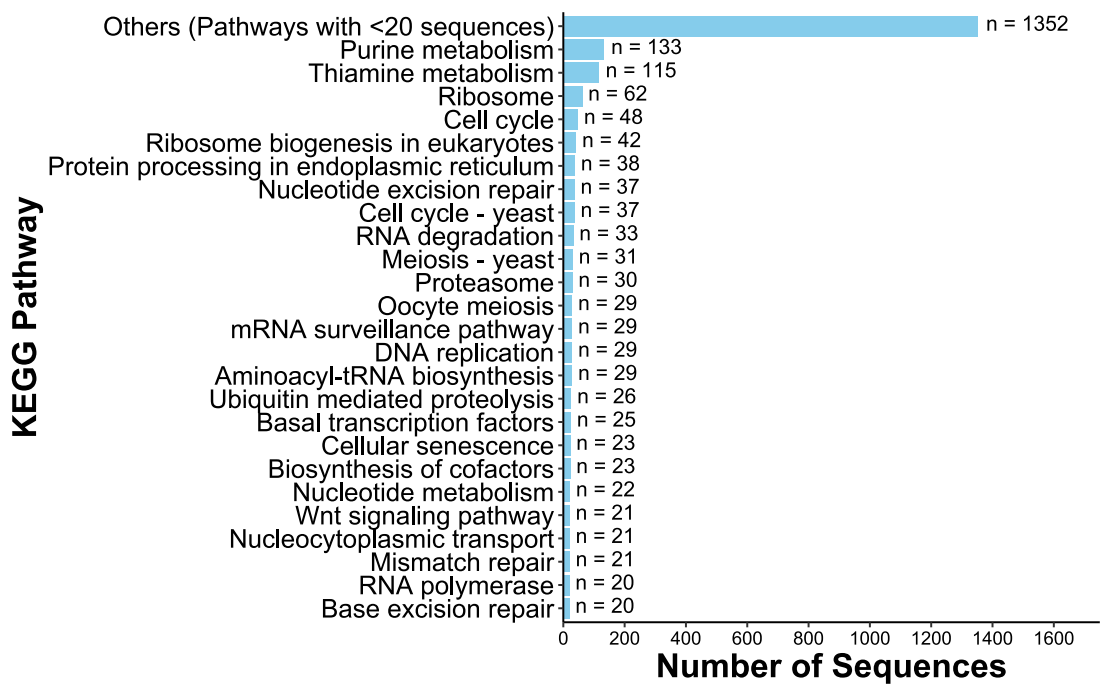


Figure 4.10: Sequence distribution in each KEGG pathway highlighting the top 25 pathways represented in the dataset. A majority of the sequences were involved in purine and thiamine metabolism, cellular processes and genetic information processing.

4.5.4 *Microsporidia* sp. MB has retained most of the genes involved in the glycolytic pathway

Microsporidia sp. MB was found to contain 10 out of the 11 proteins involved in the glycolytic pathway, as shown in Figure 4.11. Interestingly, among the Enterozooonida species, *E. cuniculi* was observed to have retained all the proteins in this pathway, while *E. bieneusi* seems to have lost all of them. This could be attributed to the fact that *E. cuniculi* has been extensively studied, sequenced, and fully annotated and is often used as a reference point for most microsporidia gene studies. On the other hand, the absence of 3-phosphoglycerate kinase in *Microsporidia* sp. MB might suggest that this gene has not been sequenced yet. Furthermore, its closest sister *V. corneae* and *Nosema granulosis* had the same set of proteins involved in the glycolytic pathway, as illustrated in Figure 4.11.

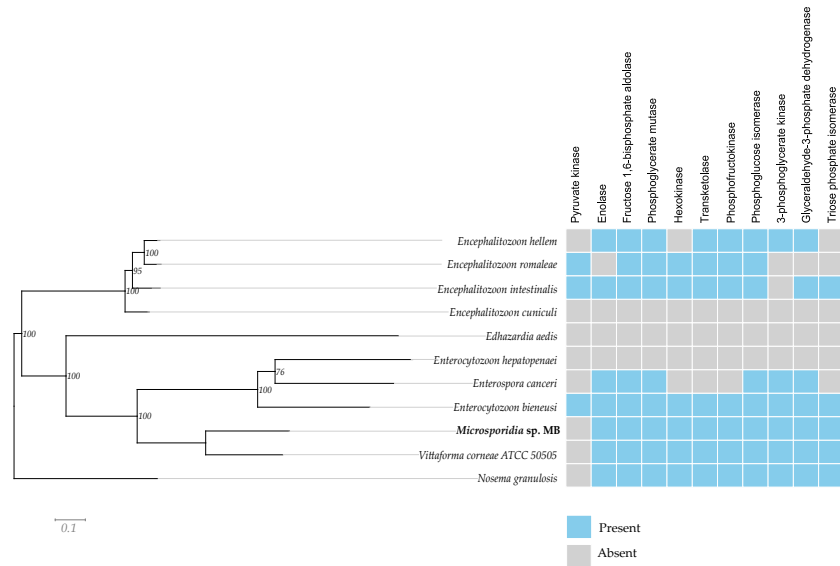


Figure 4.11: Overview of the loss and retention of key genes involved in the glycolytic pathway genes across different microsporidia species. Retained genes are highlighted in light blue in the heatmap, while those that have been lost are highlighted in gray. *Microsporidia* MB has retained most of the genes involved in glycolysis except for Pyruvate kinase, similar to *V. corneae* and *N. granulosis*. Comparatively, *E. aedis*, *E. cuniculi* and *E. hepatopenaei* appear to have lost all these genes, suggesting a total reliance on the host’s glycolysis function. The maximum likelihood phylogenomic tree on the left was generated using IQ-TREE v1.6.12 based on the 196 common BUSCO single-copy orthologous genes found in all species. Percent bootstrap support values are shown on each node and the branch length scale bare at the bottom of the tree.

4.5.5 *Microsporidia* sp. MB contains all three key components of the RNAi machinery

Orthologs of all three components of the RNAi machinery were present in *Microsporidia* sp. MB. These included the Dicer (gene_297), RNA-dependent RNA polymerase (gene_163), and Argonaute (gene_271) genes. The similarity of these orthologs to proteins found in other microsporidia species was further confirmed through phylogeny and conserved domain analyses. Low levels of TEs in this genome (highlighted in Chapter 3) could be explained by the presence of the key RNAi pathway components that have been reported to silence TE expression

[Obbard et al., 2009].

4.5.5.1 Dicer

Phylogenetic analysis of the Dicer orthologs found in 15 microsporidia species showed high similarity between *Microsporidia* sp. MB and *V. corneae* ATCC50505 (Figure 4.12). An analysis of the conserved domains of the Dicer gene on CDD revealed that this protein contained two main conserved domains, including the Ribonuclease III C terminal domain (RIBOc) and the dsRNA-specific ribonuclease domain (Rnc) (Figure 4.12b). Multiple sequence alignment showed the presence of inserts between the RIBOc and Rnc domains in *Microsporidia* sp. MB, *E. breve*, *P. epiphaga*, and *P. philotis*.

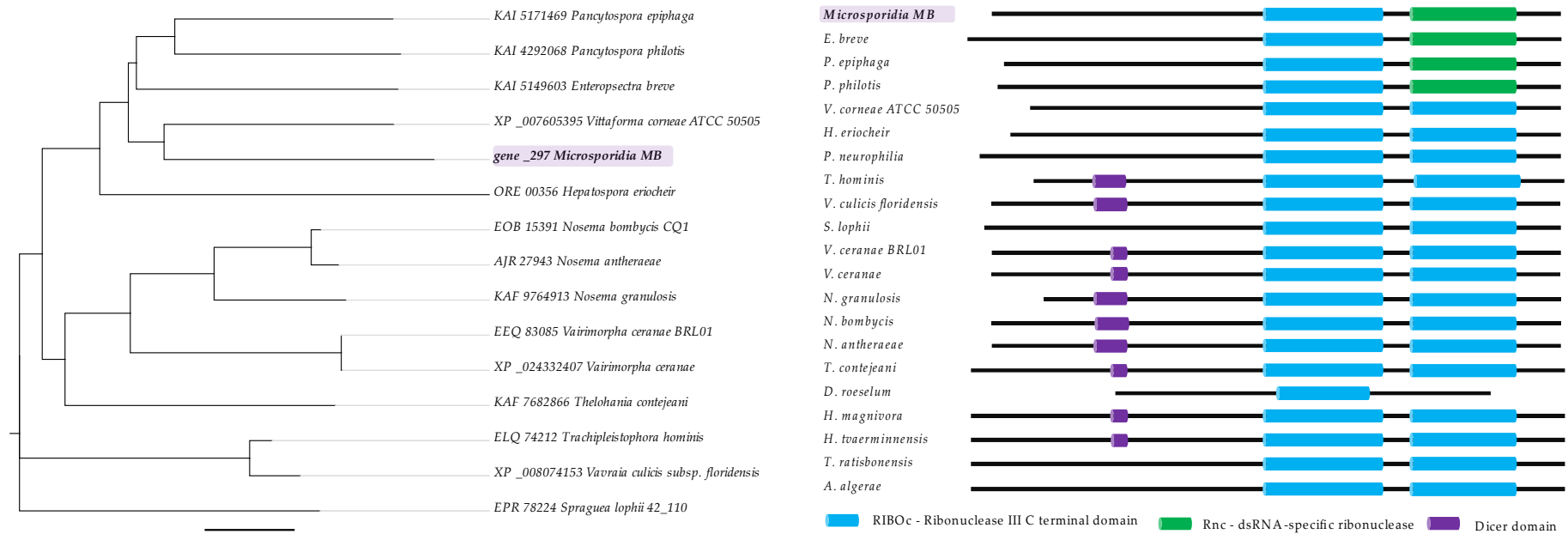


Figure 4.12: Analysis of the Dicer ortholog in *Microsporidia* sp. MB (a) A rooted phylogenetic tree of Dicer genes in various microsporidian species. The tree was constructed using The scale for the branch length is shown at the bottom. (b) A schematic representation of the position of conserved domains in the multiple sequence alignment is shown. An insert was observed between the RIBOc and Rnc domains in 4 of the species including *Microsporidia* sp. MB, *E. breve*, *P. epiphaga* and *P.philotis*

4.5.5.2 Argonaute

The central component of the RNA-induced silencing complex (RISC) and related complexes is Argonaute. This protein contains the Piwi and a portion of Argo subdomains, one of which provides the 5' anchoring of the guide RNA and the other the catalytic site for slicing. This domain is also found in closely related proteins, including the Piwi subfamily, observed in *Microsporidia* sp. MB (Figure [4.13](#)).

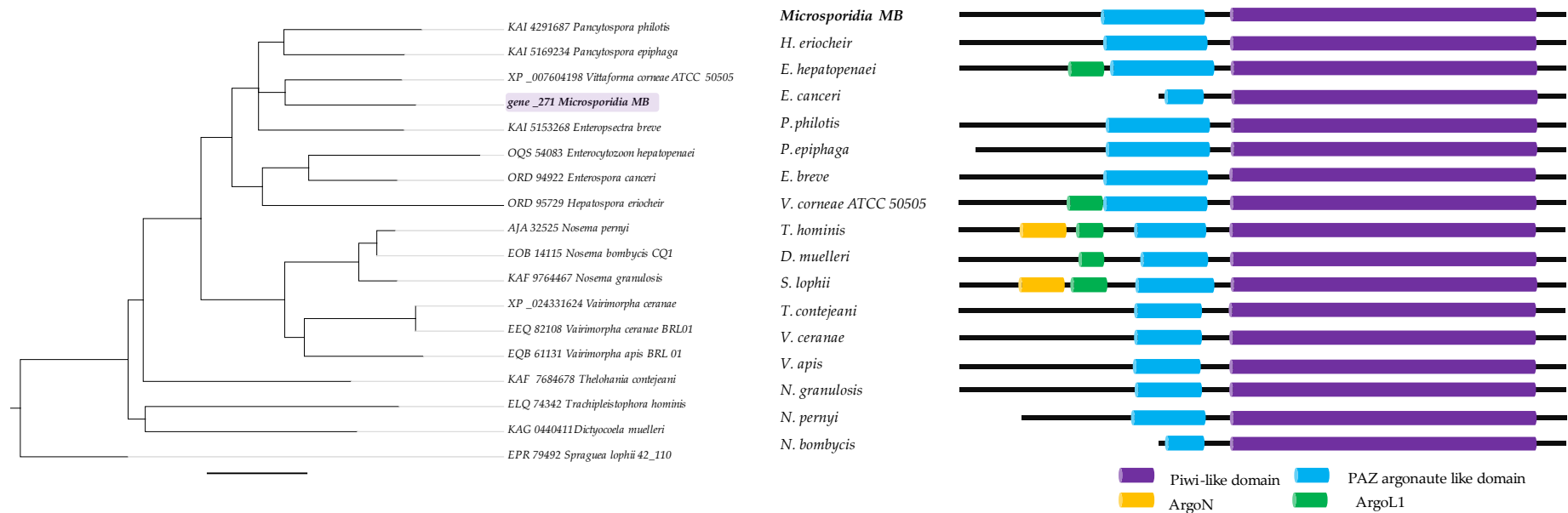


Figure 4.13: Analysis of Argonaute orthologs. (a) A rooted phylogenetic tree of Argonaute orthologs in microsporidia. The tree was constructed on protein sequences of the Argonaute orthologs. A thousand bootstraps are indicated at the nodes. The scale for the branch length is shown at the bottom. (b) A schematic representation of the position of conserved Piwi-like and PAZ domains mapped onto the multiple sequence alignment.

4.5.5.3 RNA-dependent RNA polymerase

On phylogenetic analysis, high similarity was observed between the orthologs in *V. corneae* and *Microsporidia* sp. MB for RNA-dependent RNA polymerase, which is involved in post-transcriptional gene silencing. This was further confirmed by the presence of the RdRP conserved region in this ortholog (Figure [4.14](#)).

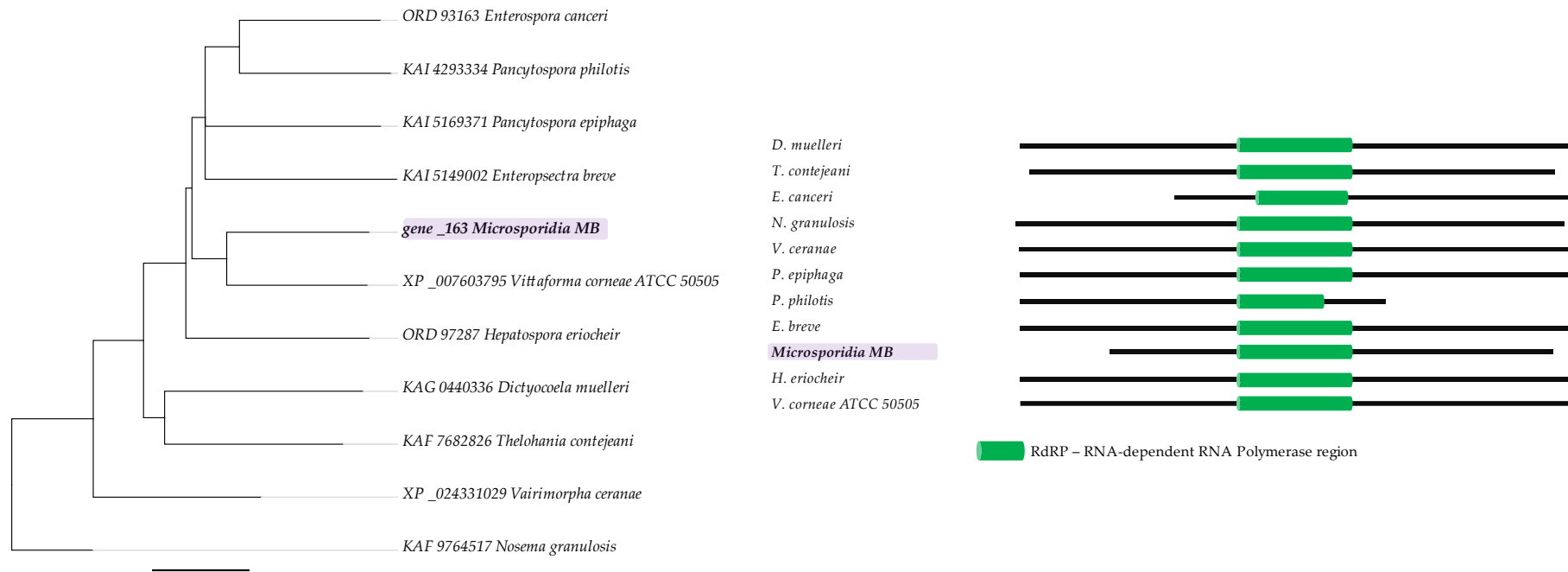


Figure 4.14: Analysis of the RNA-dependent RNA polymerase. (a) Phylogeny studies indicate the similarity between this ortholog in *Microsporidia* sp. MB and other microsporidias. (b) Identification of the conserved RdRP domain across different species.

4.5.6 The presence of several MRGs in *Mirosporidia* sp. MB suggests that it may occasionally undergo meiosis during sporogony

Recent studies have described the meiotic processes occurring in microsporidia during the early phases of sporogony [Lee et al., 2014]. The merging of haploid nuclei produces a single diploid nucleus, which divides during mitosis to create a new diplokaryon with two diploid nuclei. These nuclei then form synaptonemal complexes and fuse to form a single tetraploid nucleus, where chromosomal mingling occurs. Finally, two consecutive reductive divisions result in the production of eight haploid nuclei. Table 4.1 below highlights key proteins involved in different stages of meiosis.

Table 4.1: An overview of the different stages of meiosis and the proteins involved in each stage.

Stages of meiosis	Protein products involved
Activation of transcription of meiotic genes	Ime1
Sister chromatids join and homologs synapse	Rec8, Hop2, Mnd1, Rad21, Smc1-6
Formation of double-strand breaks in DNA	Spo11, Rec104
Formation of the synaptonemal complex	Red1, Hop1, Zip1 & 2
Resection of DNA ends and attachment of inter-homolog recombination proteins	Rec114, Rad1, Rad50, Mre11
Homology searching and strand exchange	Dmc1, Rad51, Rad52, Rad55, Rad57
Formation of stable heteroduplex, strand invasion, and DNA synthesis Mismatch and frameshift repair	Mlh1-3, Msh2, Msh6, Pms1
Mediation of crossover interference and resolution of Holliday junctions	Msh4, Msh5, Mer3 (HFM1)
Chromosome segregation	Ubc11, Mam1

Despite the reduced complexity of microsporidia genomes, investigations have revealed genes that show homology or similarity to known meiotic genes in other eukaryotes [Ndikumana et al., 2017]. These genes include homologs of meiosis-specific genes identified in other organisms, such as Dmc1, Msh4, and Msh5, as detailed in Appendix Table A3. These homologs are integral to fundamental meiotic processes, including genetic recombination, chromosomal synapsis, chromosome segregation, and spore formation. This indicates that microsporidia have retained critical genetic components required for meiosis despite their significantly reduced genomes, albeit with certain modifications and adaptations to their parasitic lifestyle and evolutionary history. Eighteen MRG orthologs were identified in the genome (Table A3). The presence of these meiotic genes orthologs in the *Microsporidia* sp. MB genome suggests its potential mode of reproduction. These genes aid in the study of the reproductive strategies and evolutionary biology of the symbiont.

4.5.6.1 Role of the SMC proteins in meiosis

SMC proteins are involved in critical processes for accurately segregating proteins during meiosis. These processes include chromosome condensation, sister-chromatid cohesion, and DNA repair and recombination [Strunnikov, 1998, Strunnikov and Jessberger, 1999, Jessberger, 2002]. Five SMC orthologs (SMC1, 2, 3, 4, and 6) were identified in *Microsporidia* sp. MB (Table A3). Their respective multiple sequence alignments and tree calculations are highlighted below.

4.5.6.1.1 SMC1 [gene_603]

SMC1 is responsible for holding sister chromatids together during both mitosis and meiosis in the cohesion complex [Bolaños-Villegas, 2021]. It has also been implicated in meiotic recombination regulation as a heterodimer with SMC3 [Bolaños-Villegas, 2021]. The putative gene, gene_603, in the *Microsporidia* sp. MB genome showed high similarity to SMC1 of *Enteropsectra breve* and *Pancytospora epiphaga* (Figure 4.15)

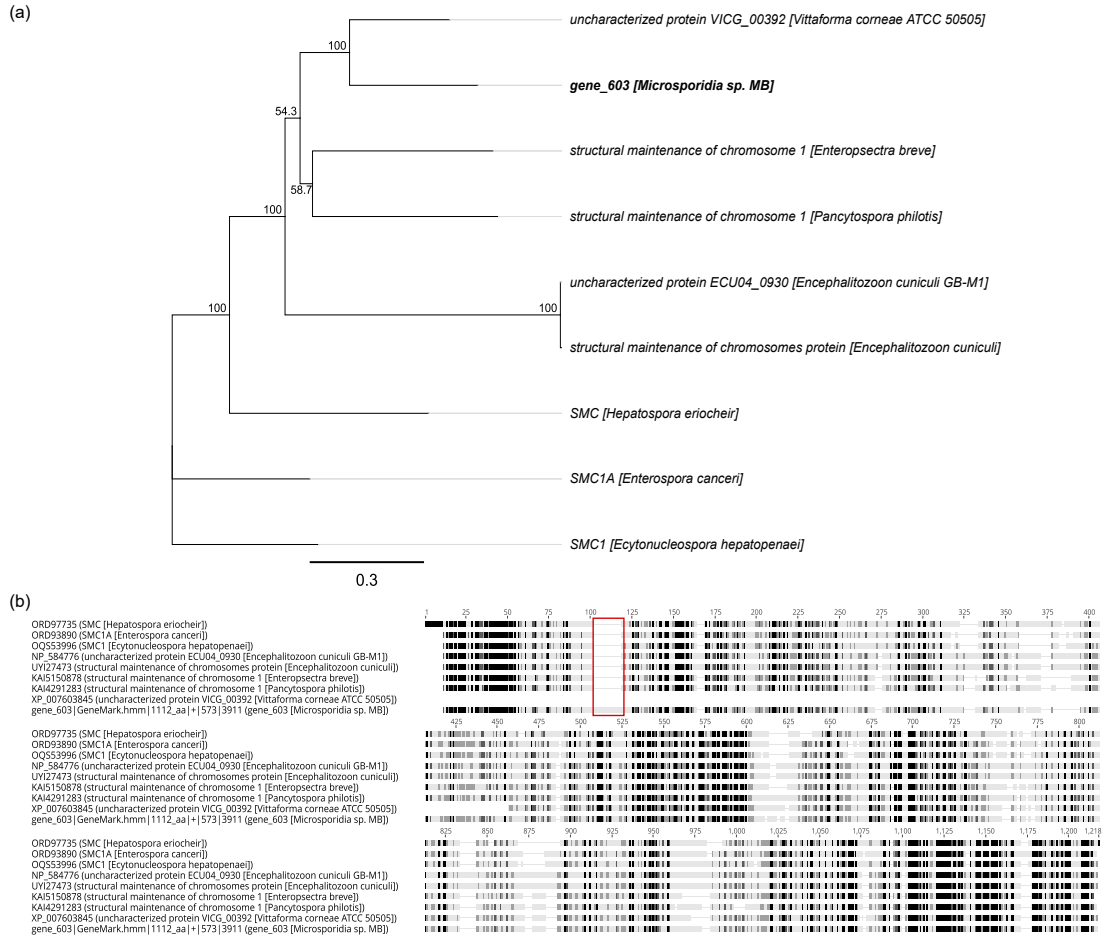


Figure 4.15: Overview of the SMC1 ortholog in *Microsporidia* sp. MB.

4.5.6.1.2 SMC2 [gene_89]

SMC2, a condensin complex component, is crucial in condensing chromosomes into tightly packed structures during meiosis. This process is essential for adequately segregating chromosomes [Yokomori, 2003]. To achieve this, the condensin complex utilizes ATP energy to create positive supercoiling in DNA, resulting in the formation of the mitotic chromosome scaffold [Pavlova and Zakian, 2003].

Gene_89 (Figure 4.16) in the symbiont was identified as an SMC2 protein. It similarly showed a close relationship with the characterized SMC6 proteins of *Enteropsectra breve* and *Pancytophora epiphaga*, and contained an insert at residue 823-845 (Figure 4.16b).

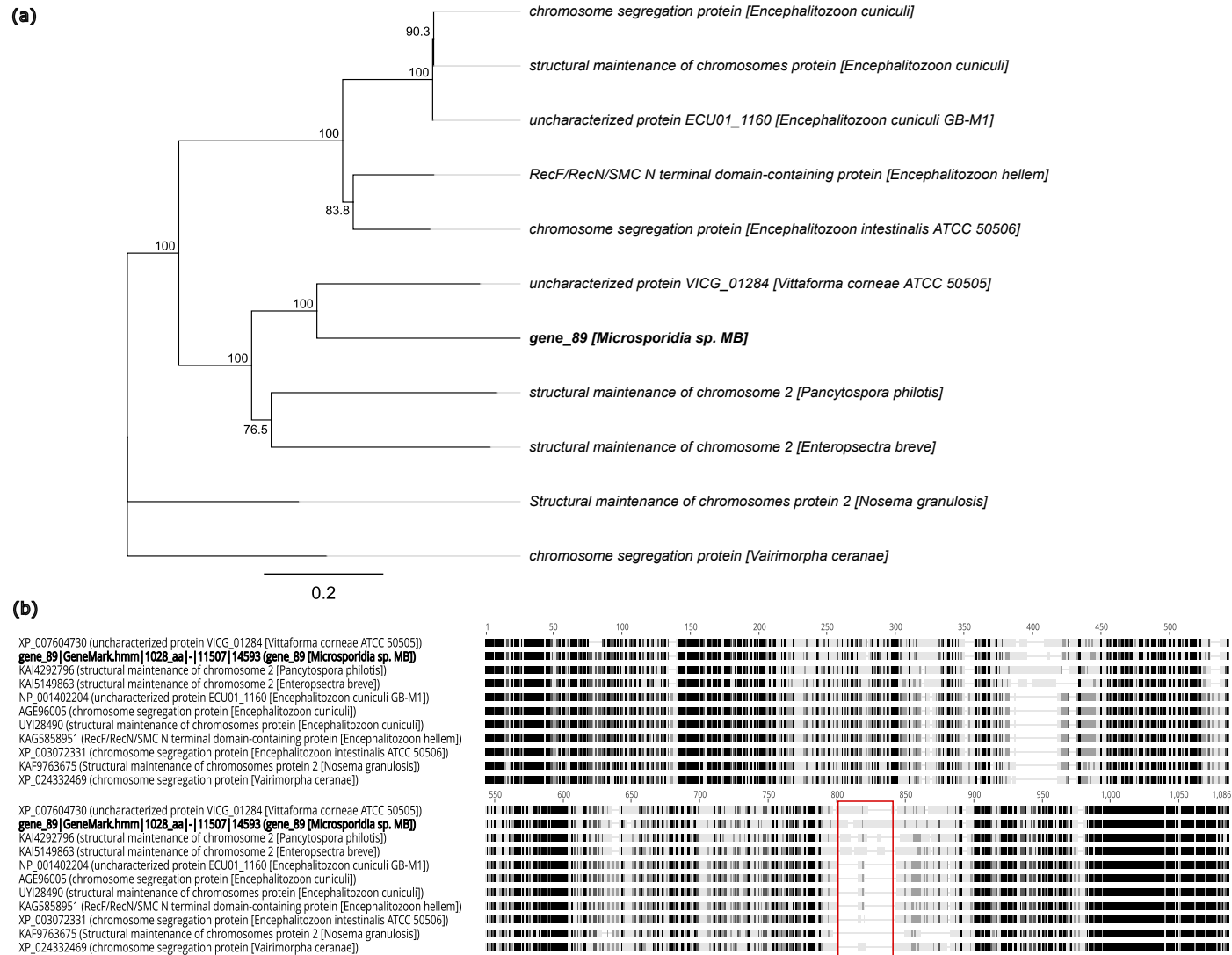


Figure 4.16: Analysis of the gene_89 SMC2 ortholog.

4.5.6.1.3 SMC3 [gene_1347]

The SMC3 protein forms a heterodimer with SMC1 in the cohesin complex to regulate meiosis recombination [Bolaños-Villegas, 2021]. The SMC3 homolog in *Microsporidia* sp. MB was identified as gene_1347 similar to *Pancytospora* SMC3 proteins (Figure 4.17).

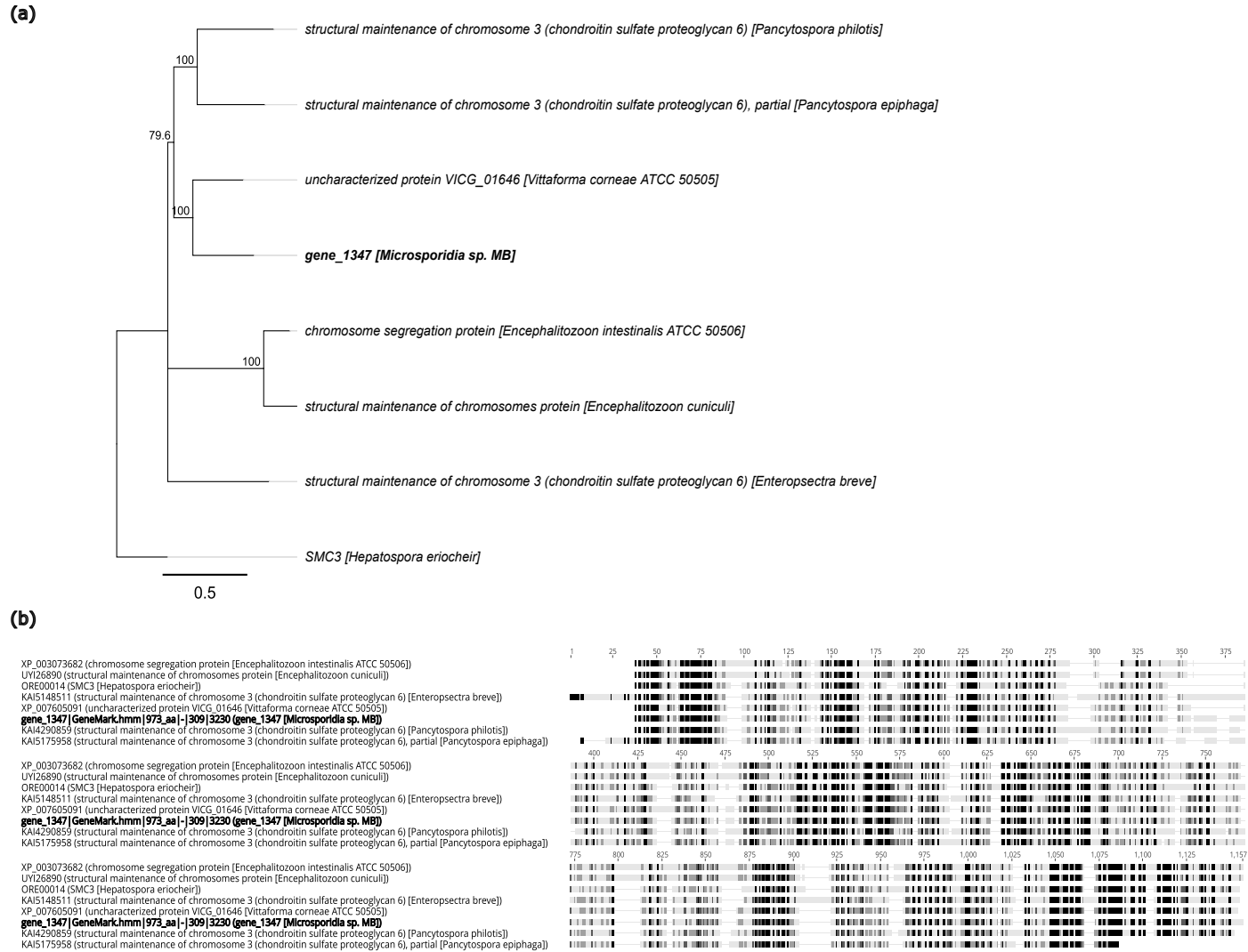


Figure 4.17: The SMC3 ortholog in *Microsporidia* sp. MB.

4.5.6.1.4 SMC4 [gene_1051]

SMC4 is involved in the structural reorganization of meiotic chromosomes [Ball Jr and Yokomori, 2001]. It forms the condensin complex with SMC2 and is involved in chromosome condensation. This complex uses ATP during chromosome compaction, forming mitotic chromosome scaffolds [Yokomori, 2003]. Gene_1051 was closely associated with *Enteropsectra breve* and *Pancytospora epiphaga* (Figure 4.18).

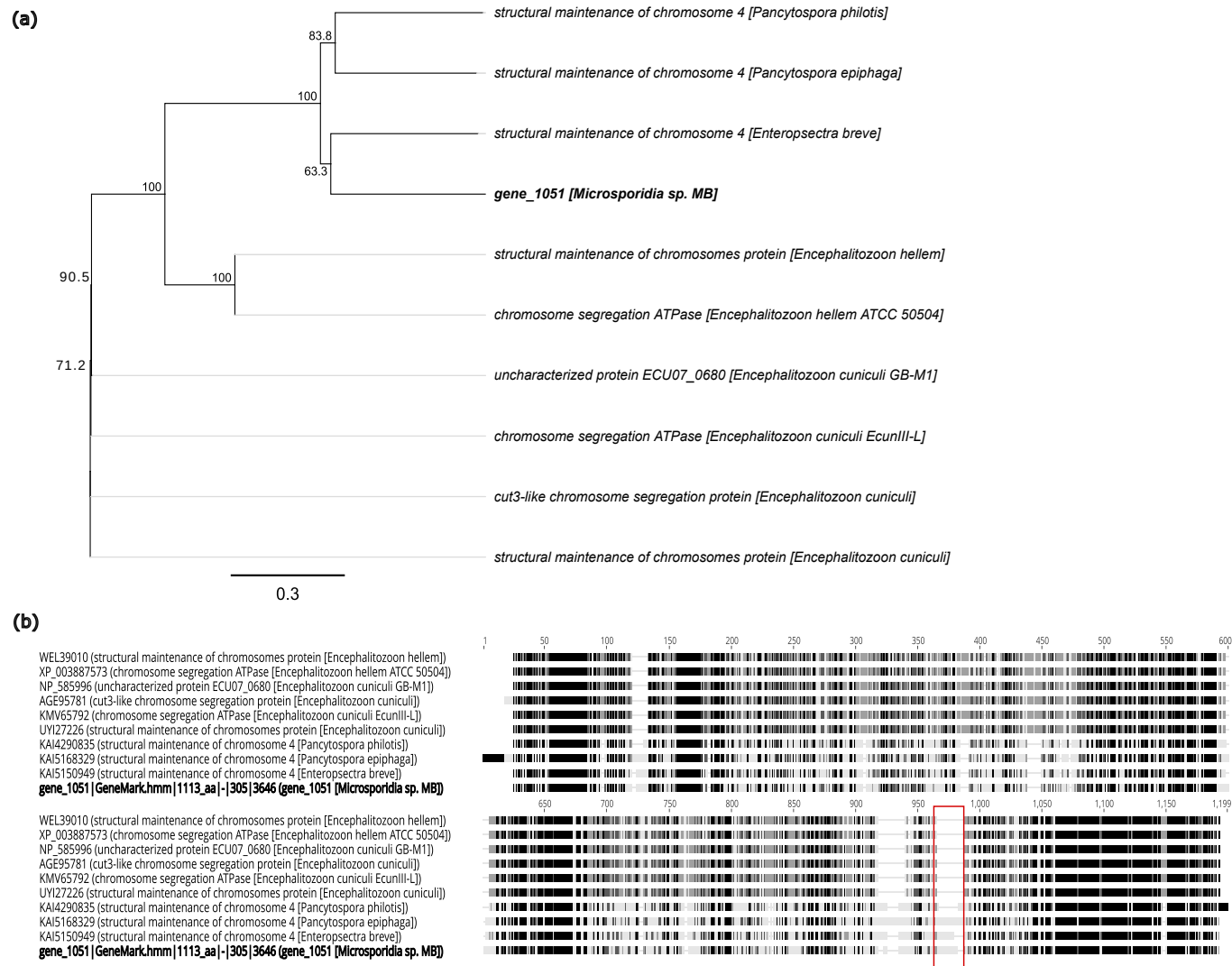


Figure 4.18: Analysis of the similarity between the SMC4 orthologs in *Microsporidia* sp. MB and other species.

4.5.6.1.5 SMC6 [gene_941]

The SMC5/6 complex is involved in the repair of programmed DNA double-strand breaks (DSBs) through homologous recombination during meiosis [Bickel et al., 2010]. Mutants of *Caenorhabditis elegans* that lack functional SMC-5/6 proteins exhibit chromosome fragmentation, indicating the critical role played by this complex in the successful completion of meiosis [Bickel et al., 2010]. In *Microsporidia* sp. MB, one homolog of the SMC5/6 mutant, was identified (gene_941), showing close similarity to SMC6 of *Enteropsectra breve* and *Pancytospora epiphaga* (Figure 4.19).

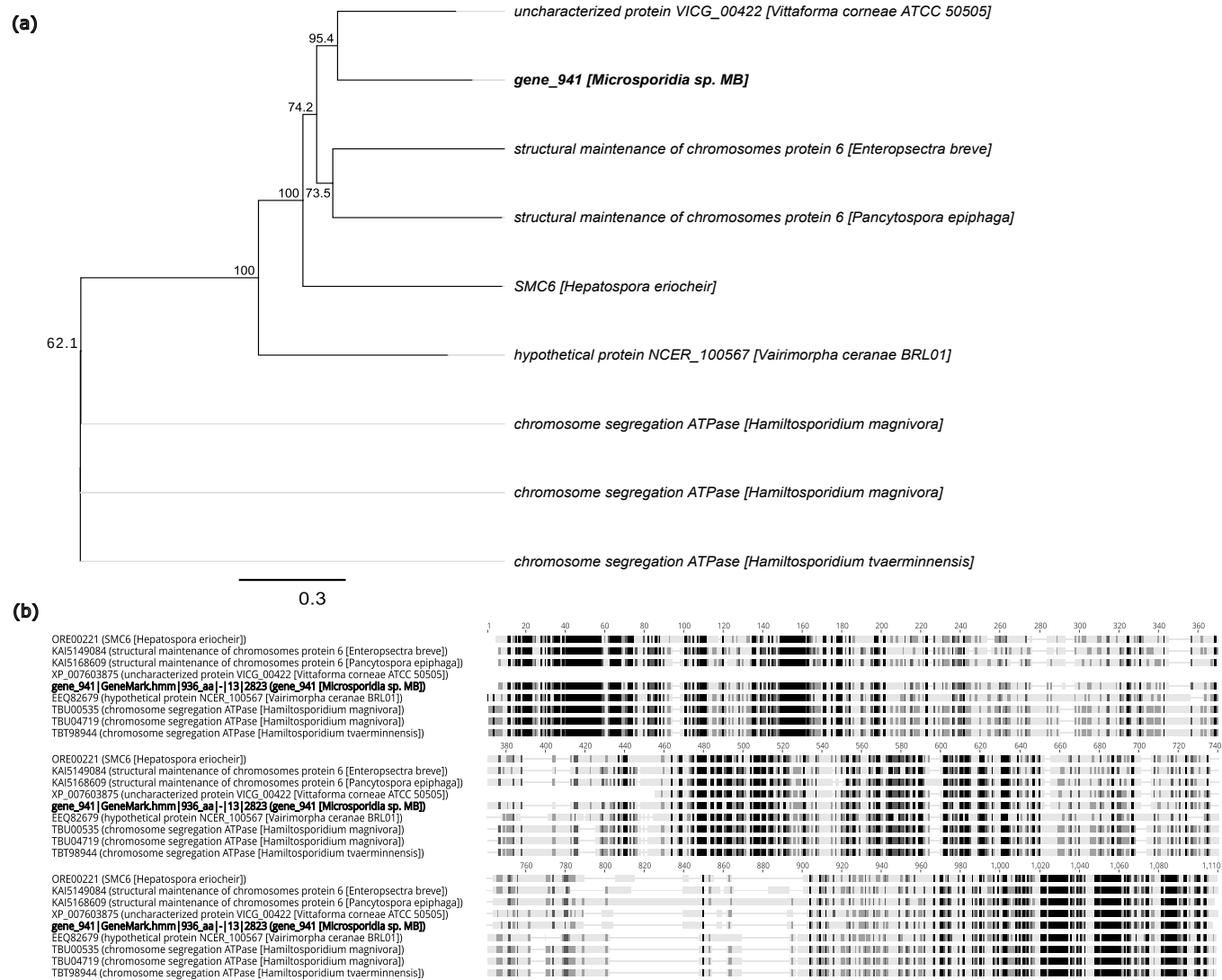


Figure 4.19: A highlight of SMC6 orthologs in microsporidia.

4.5.6.2 Sister chromatid cohesion 1 protein 3 (Scc3) [gene_1754]

Scc3 is involved in ensuring accurate chromosome segregation but facilitating sister chromatid cohesion. It also has other essential functions in gene regulation and chromosomal processes to ensure the successful progression of meiosis [Lin et al., 2011]. The similarity between gene_1754 and its close homologs is highlighted in Figure 4.20.

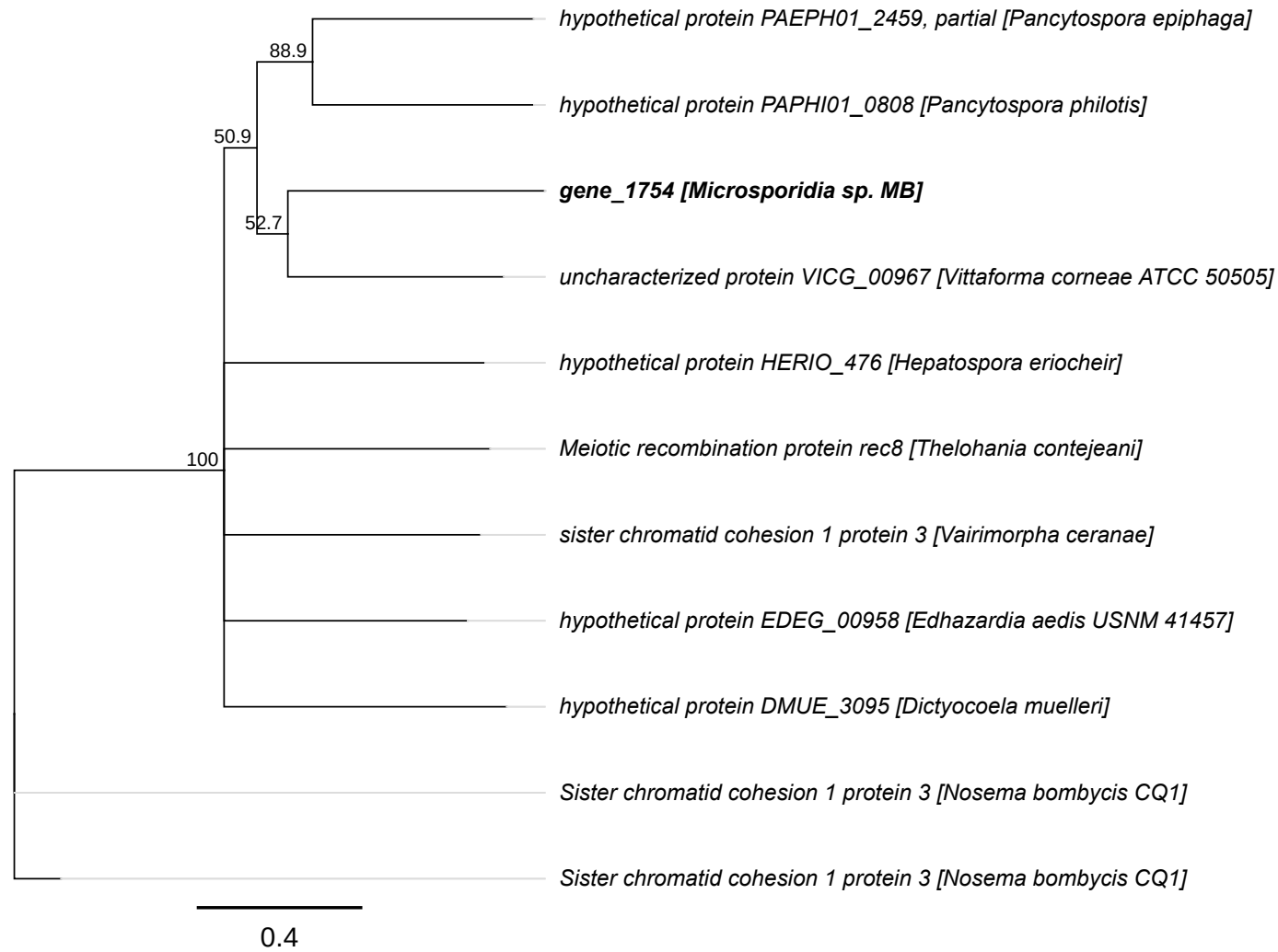


Figure 4.20: Phylogenetic comparison and msa of Sister chromatid cohesion 1 protein 3.

4.5.6.3 DNA repair and recombination proteins in meiosis

In *S. cerevisiae*, RAD51, and RAD52 proteins function in DNA repair and recombination processes during meiosis. RAD51 operates similarly to bacterial RecA and is involved in the homologous recombination pathway, where it interacts with the meiosis-specific Dmc1 protein [Shinohara et al., 1992]. RAD52 is involved in meiotic DNA metabolism and DNA damage repair [Resnick et al., 1981]. Although not extensively studied, RAD16 has been reported to participate in nucleotide excision repair (NER) in yeast [Resnick et al., 1981].

4.5.6.3.1 RAD16 [gene_1760]

Putative gene_1760 from *Microsporidia* sp. MB was closely associated with RAD16 from *Pseudoloma neurophilia* (Figure 4.21). *P. neurophilia* contained insertions at residues 223-260 and 535-550, unlike the other microsporidia species (highlighted in a red box in Figure 4.21).

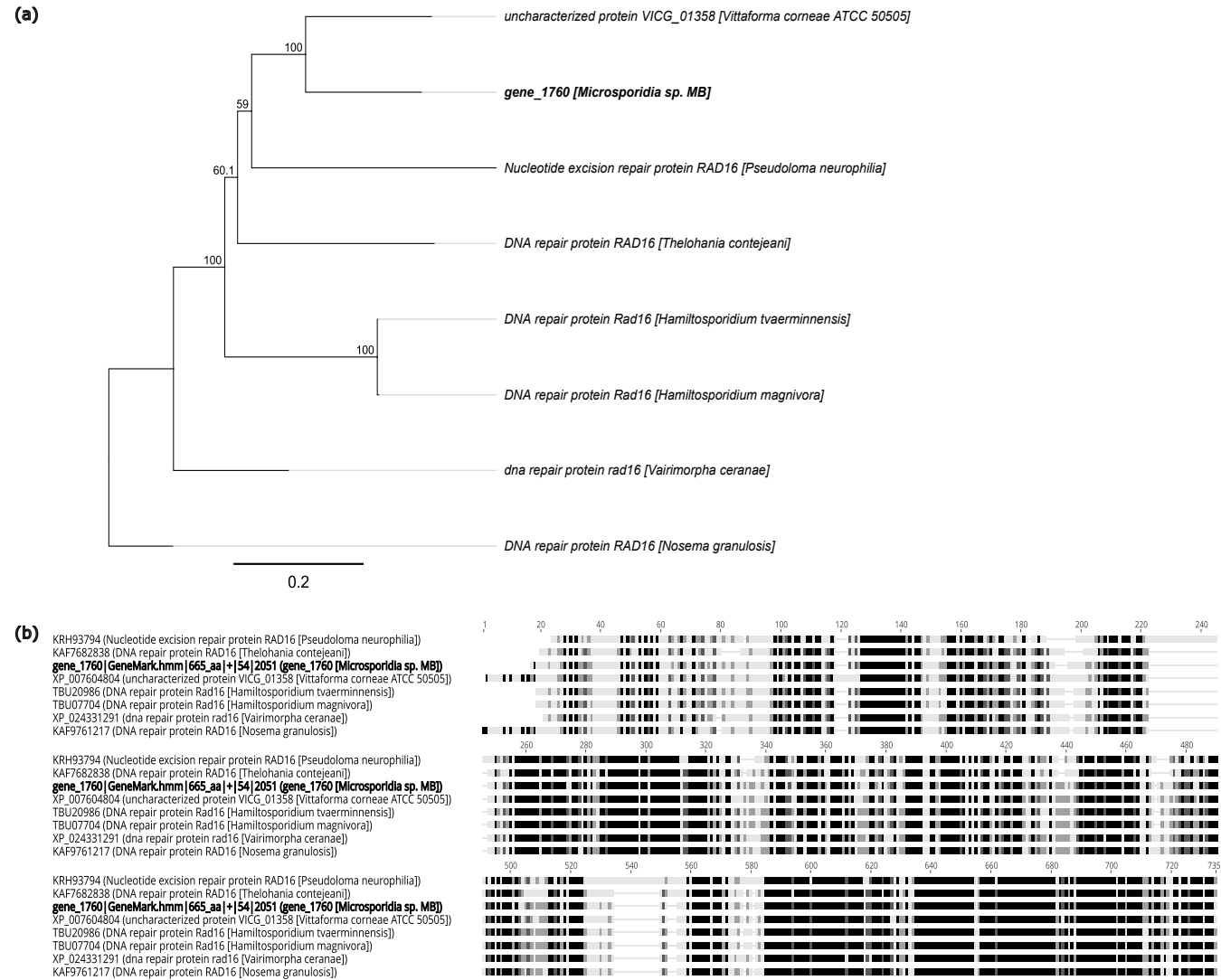


Figure 4.21: The RAD16 DNA repair and recombination protein in microsporidia.

4.5.6.3.2 RAD51 [gene_1069]

Gene_1069 had high similarity with RAD51 in *V.corneae* (Figure [4.22](#)).

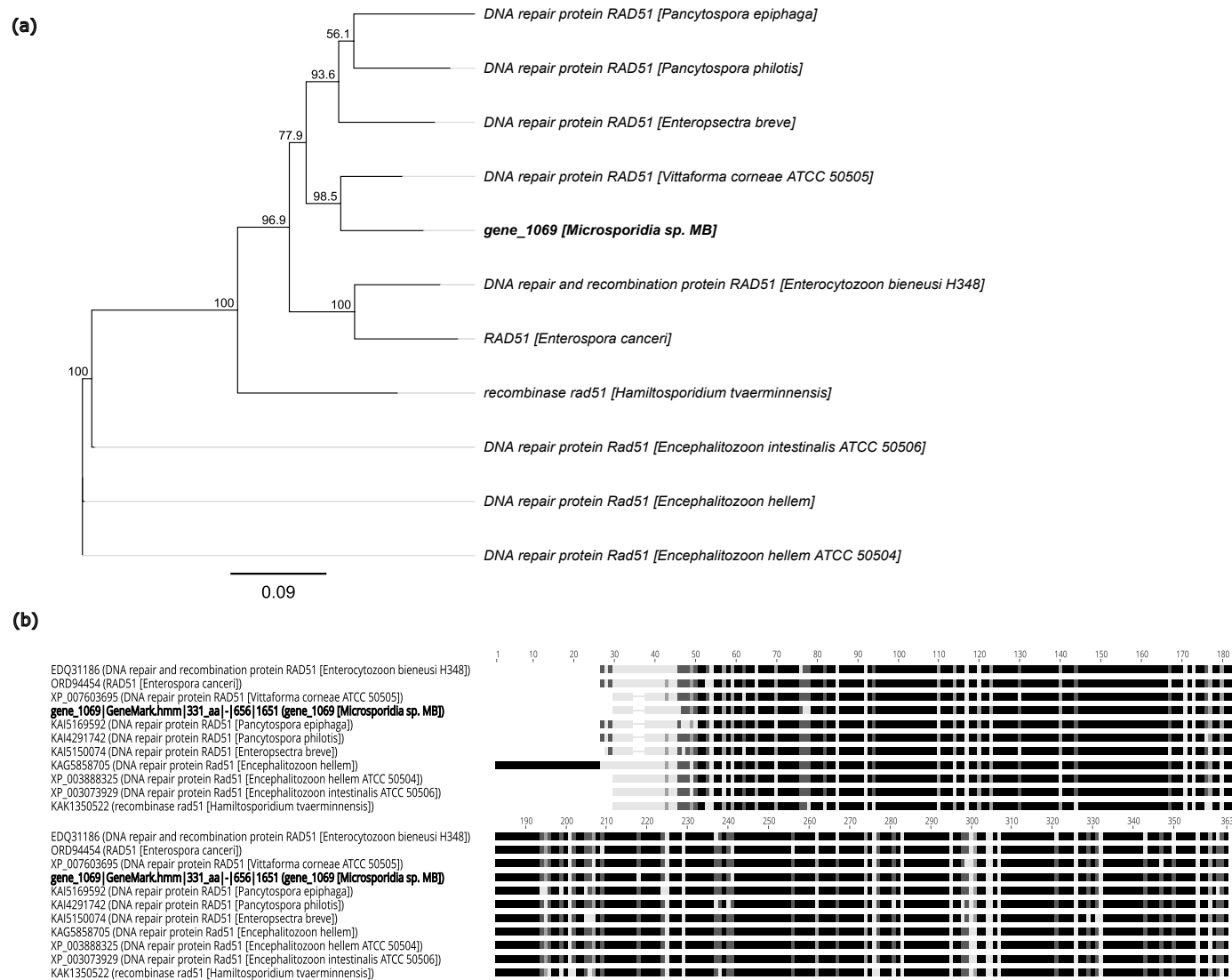


Figure 4.22: Analysis of the relationship between RAD51 orthologs in microsporidia.

4.5.6.3.3 RAD52 [gene_421]

V. corneae RAD52 protein and *Microsporidia* sp. MB had close associations (Figure 4.23)

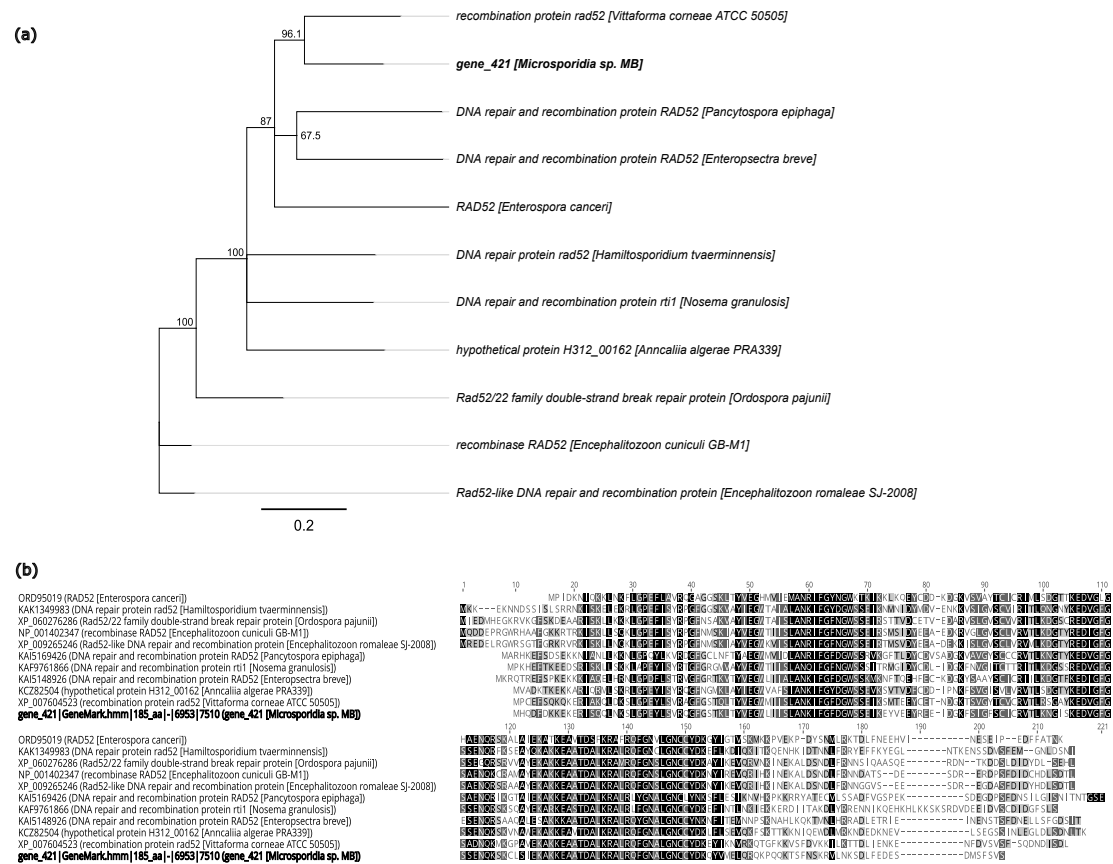


Figure 4.23: Analysis of RAD52 in microsporidia.

4.5.6.3.4 SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1/SNF2/RAD54[[gene_1398](#)]

Figure [4.24](#) highlights the presence of the in *Microsporidia* *sp. MB* ([gene_1398](#)). [[Venkataramanan et al., 2016](#)].

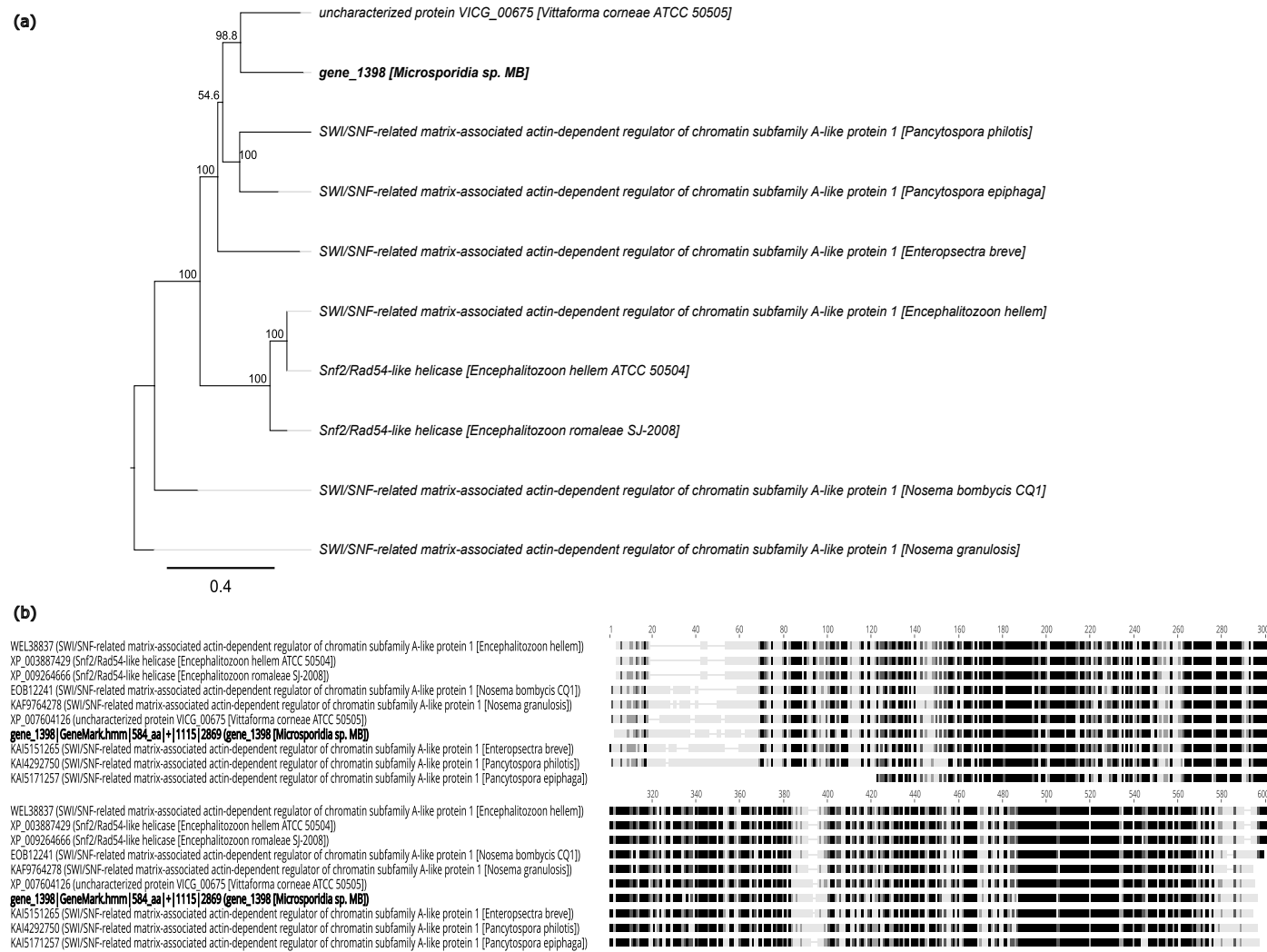


Figure 4.24: Analysis of the SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1/SNF2/RAD54.

4.5.6.4 DNA mismatch repair (MMR) proteins in meiosis

The DNA MMR complex consists of MutS, MutL, MSH2, and PMS2 proteins involved in maintaining genetic stability during meiosis [Santucci-Darmanin et al., 2002]. MutS and MutL homologs participate in post-replication DNA repair and meiotic recombination, an essential step for accurate chromosome segregation [Liu et al., 2018]. The MutS homologs, such as MSH2, recognize mismatched nucleotides and can also form complexes, such as MSH4-MSH5, which are necessary for chromosome synapsis and the crossing-over process in meiosis [Santucci-Darmanin and Paquis-Flucklinger, 2003]. MutL homologs, including PMS2, mediate multiple interactions associated with downstream MMR events [Santucci-Darmanin and Paquis-Flucklinger, 2003]. The MMR system, including MSH2 and PMS2, also has independent roles in meiosis and is implicated in repairing somatic double-strand breaks [Pannafino and Alani, 2021].

4.5.6.4.1 MutL/MLH1 [gene_478]

Figure 4.25 highlights the similarity between gene_478 in *Microsporidia* sp. MB and its counterparts. A larger insert was observed in *Microsporidia* sp. MB (residues 555-710) (Figure 4.25b).

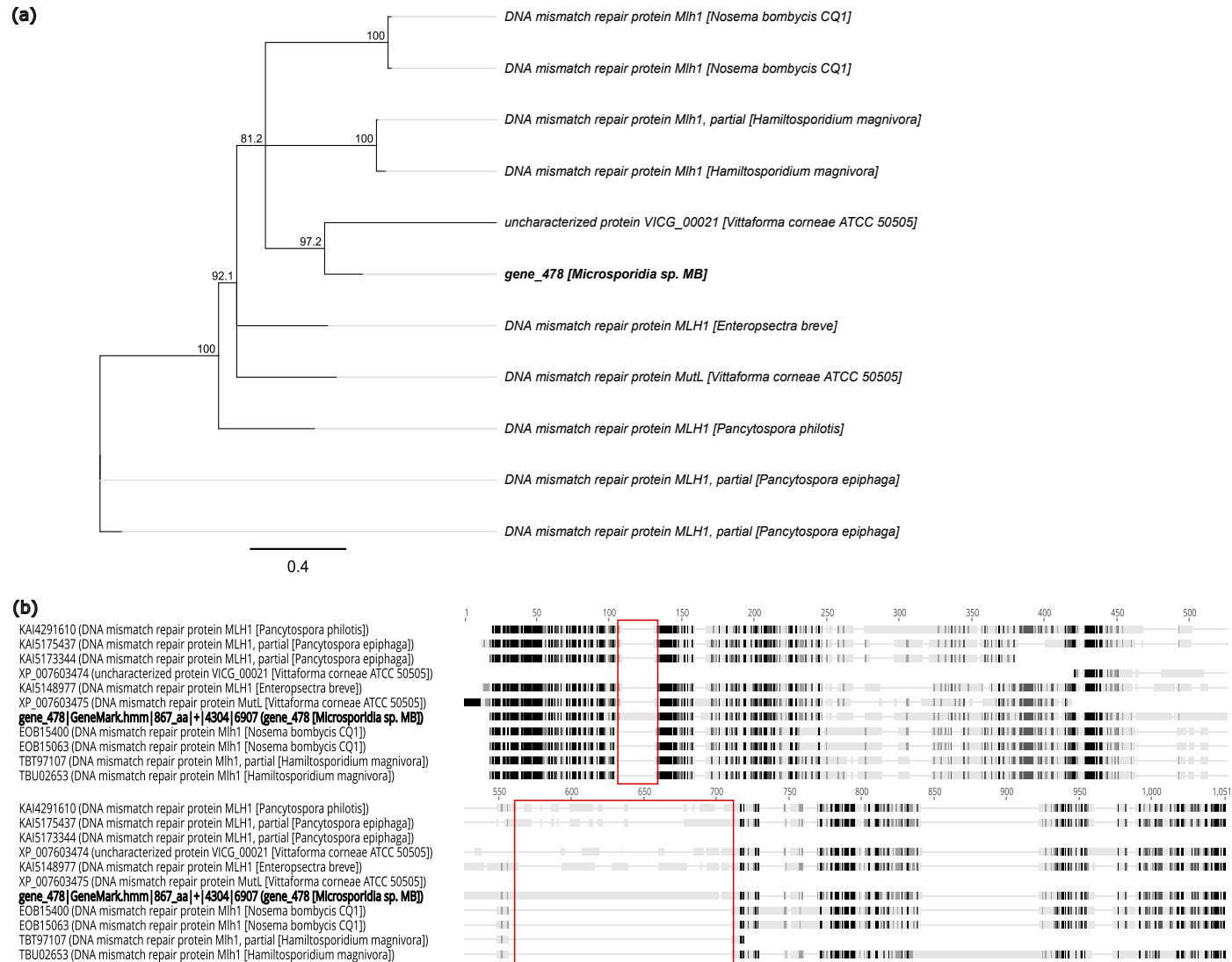


Figure 4.25: Tree calculations and sequence analysis of the MutL/MLH1 gene.

4.5.6.4.2 Msh2/MutS [gene_1101]

Associations between Msh2 from other microsporidia and gene_1101 from *Microsporidia* sp. MB is shown in Figure [4.26](#). Gene_1101 contained an insert between residue 515 and 550.

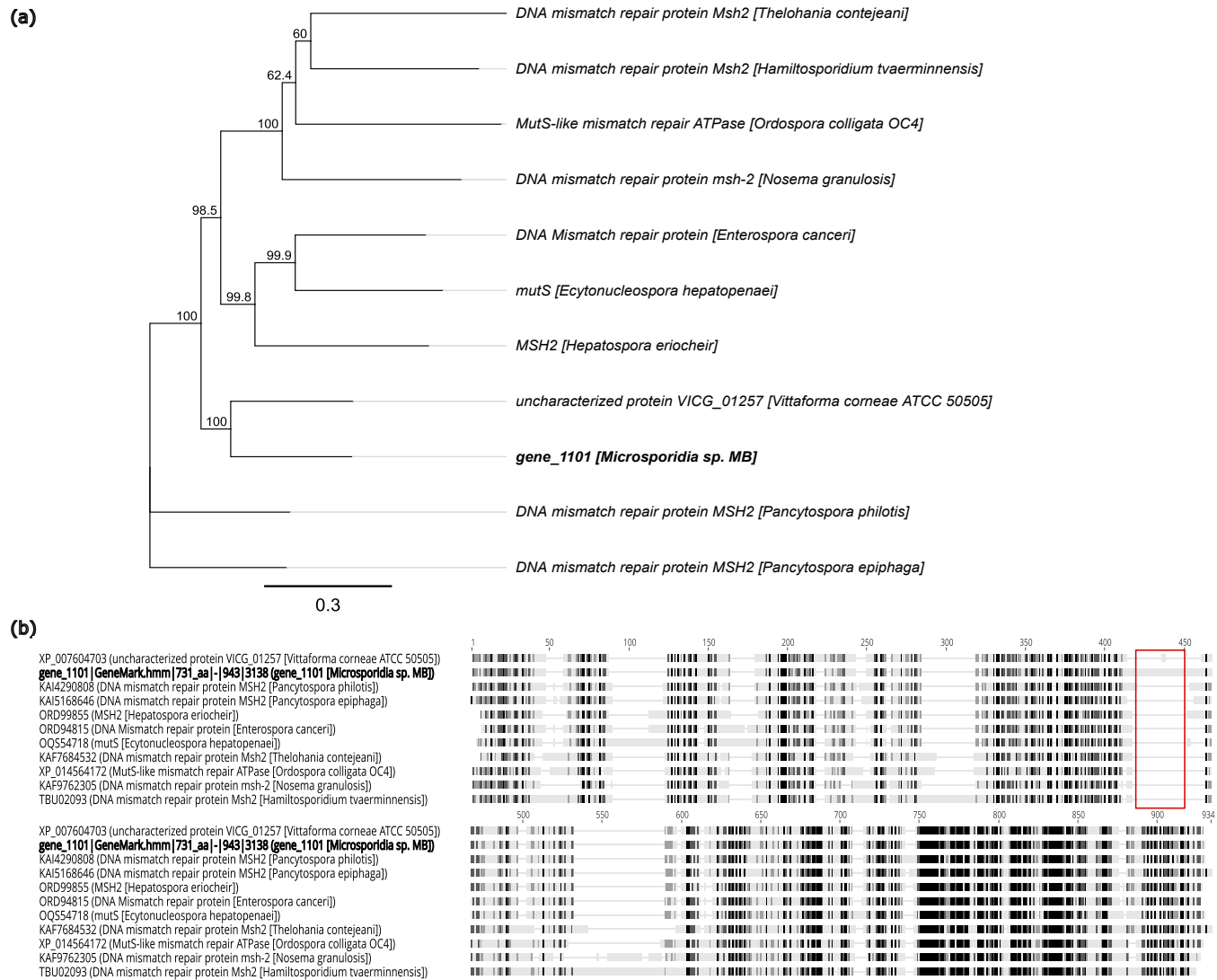


Figure 4.26: Tree calculations and sequence analysis of the Msh2/MutS gene orthologs in microsporidia

4.5.6.4.3 MutS/MSH5 [gene__684]

Figure [4.27](#) highlights the relationship between gene__684 and its homologs.

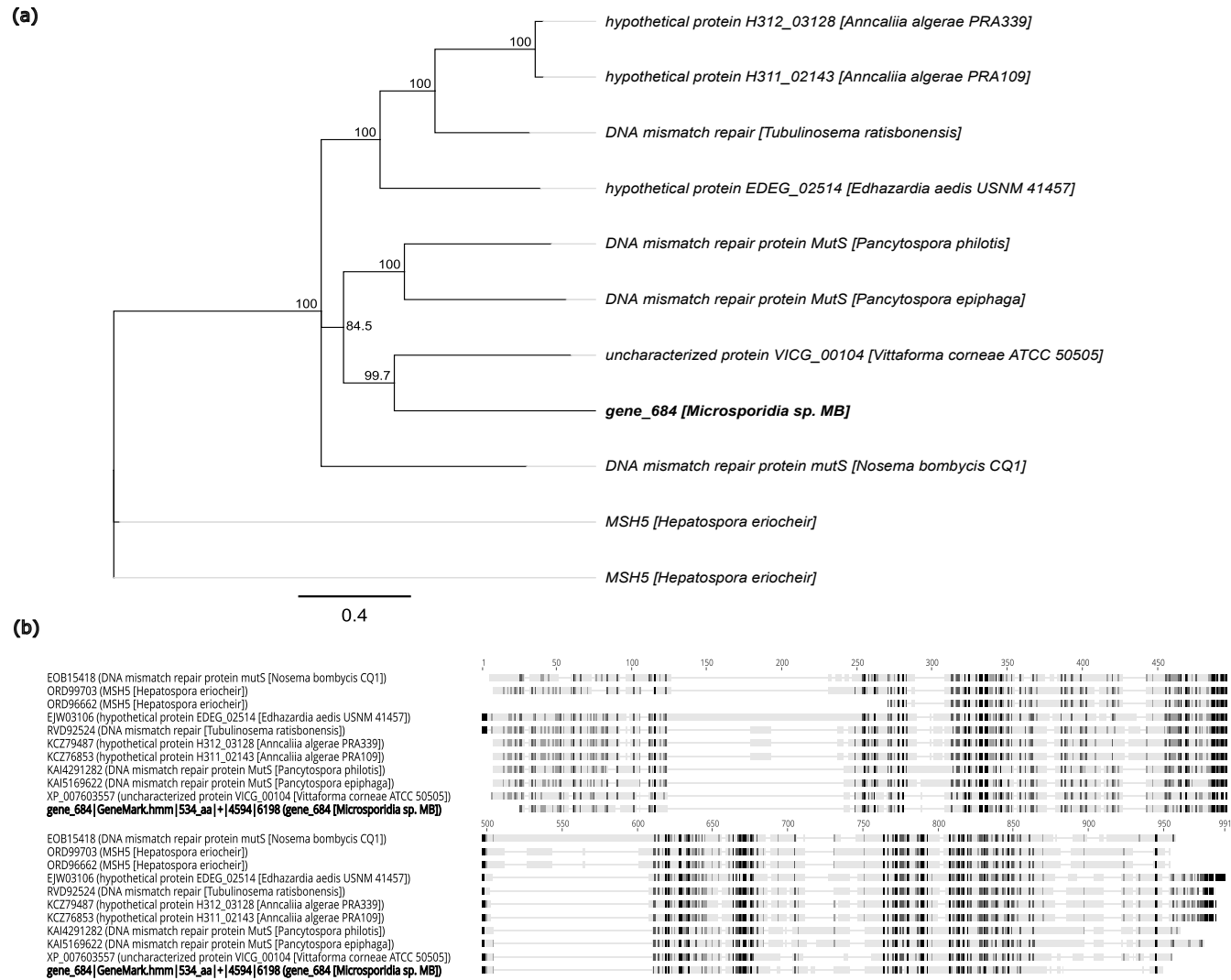


Figure 4.27: Tree calculations and sequence analysis of the MutS/MSH5 gene.

4.5.6.4.4 PMS2/MutL [gene__1355]

Microsporidia sp. MB putative gene__1355 is a PMS2 homolog (Figure [4.28](#)).

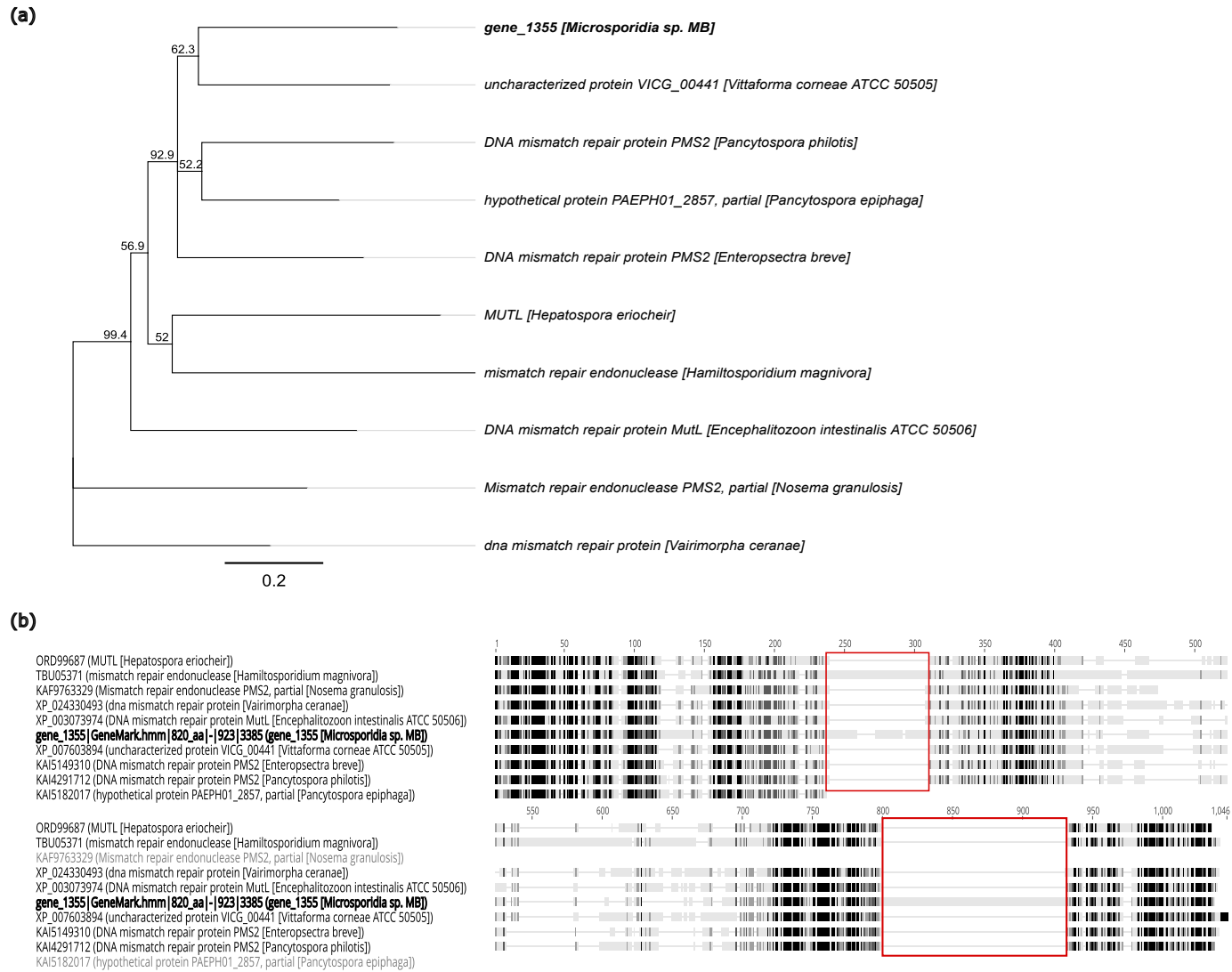


Figure 4.28: Tree calculations and sequence analysis of the PMS2/Mut gene.

4.5.6.4.5 MSH2/6 [gene__114]

Figure [4.29](#) highlights the relationship between gene__114 in *Microsporidia* sp. MB and other species.

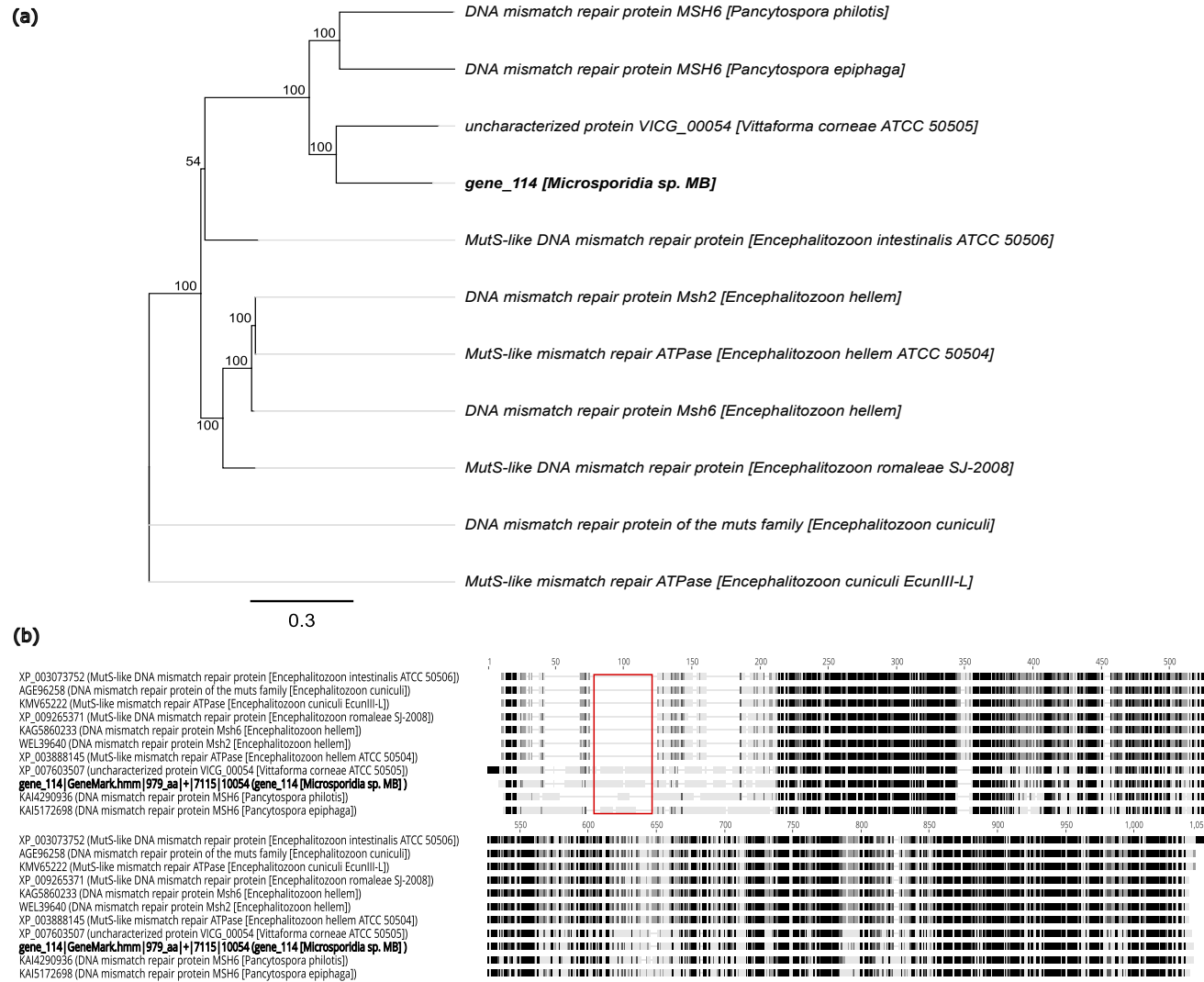
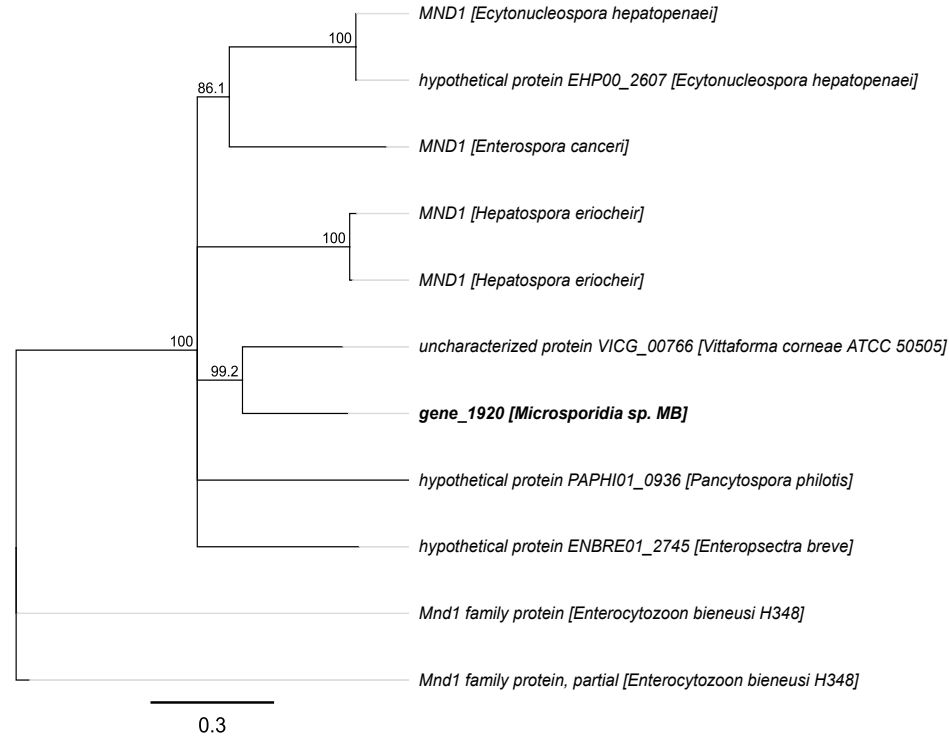


Figure 4.29: Tree calculations and sequence analysis of the MSH2/6 gene.

4.5.6.5 MND1 family protein [gene_1920]

During meiosis, the proper pairing of homologous chromosomes and the efficient repair of DSBs are crucial for forming viable gametes. The MND1 family proteins work with Hop2 in the Mnd1-Hop2 complex that interacts with recombination enzymes to facilitate homologous recombination and ensure accurate chromosome segregation by promoting proper chromosome pairing. This complex is essential for the successful formation of viable gametes [Tsubouchi and Roeder, 2002]. *Microsporidia* sp. MB contained this gene (gene_1920) (Figure 4.30).

(a)



(b)

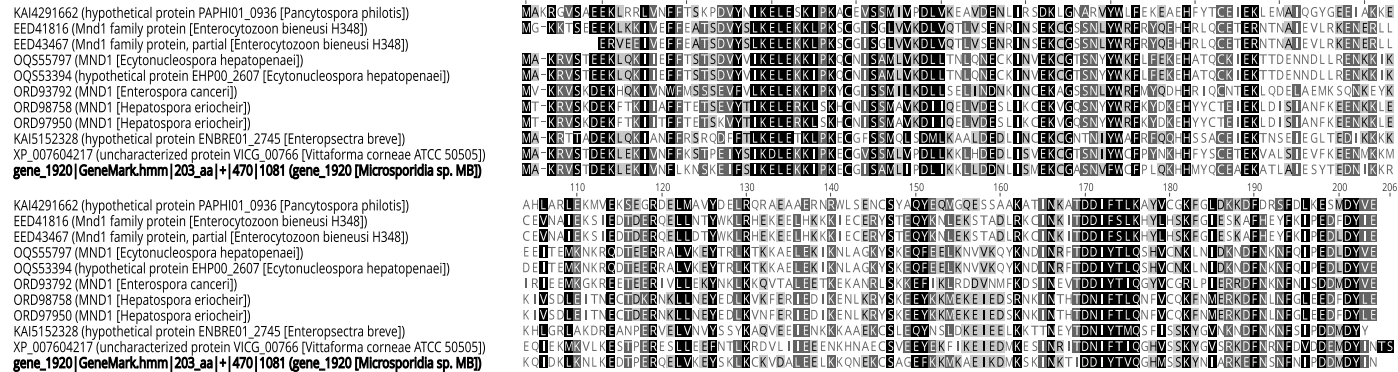


Figure 4.30: Tree calculations and sequence analysis of the MND1 family protein.

4.5.6.6 Sporulation protein RMD1 [gene_1498]

Sporulation proteins are involved in multiple stages of meiosis. The sporulation protein is involved in the production of spores during the sporogony stage of meiosis [Clancy, 1998]. Figure 4.31 highlights gene_1498 from *Microsporidia* sp. MB as a sporulation protein.

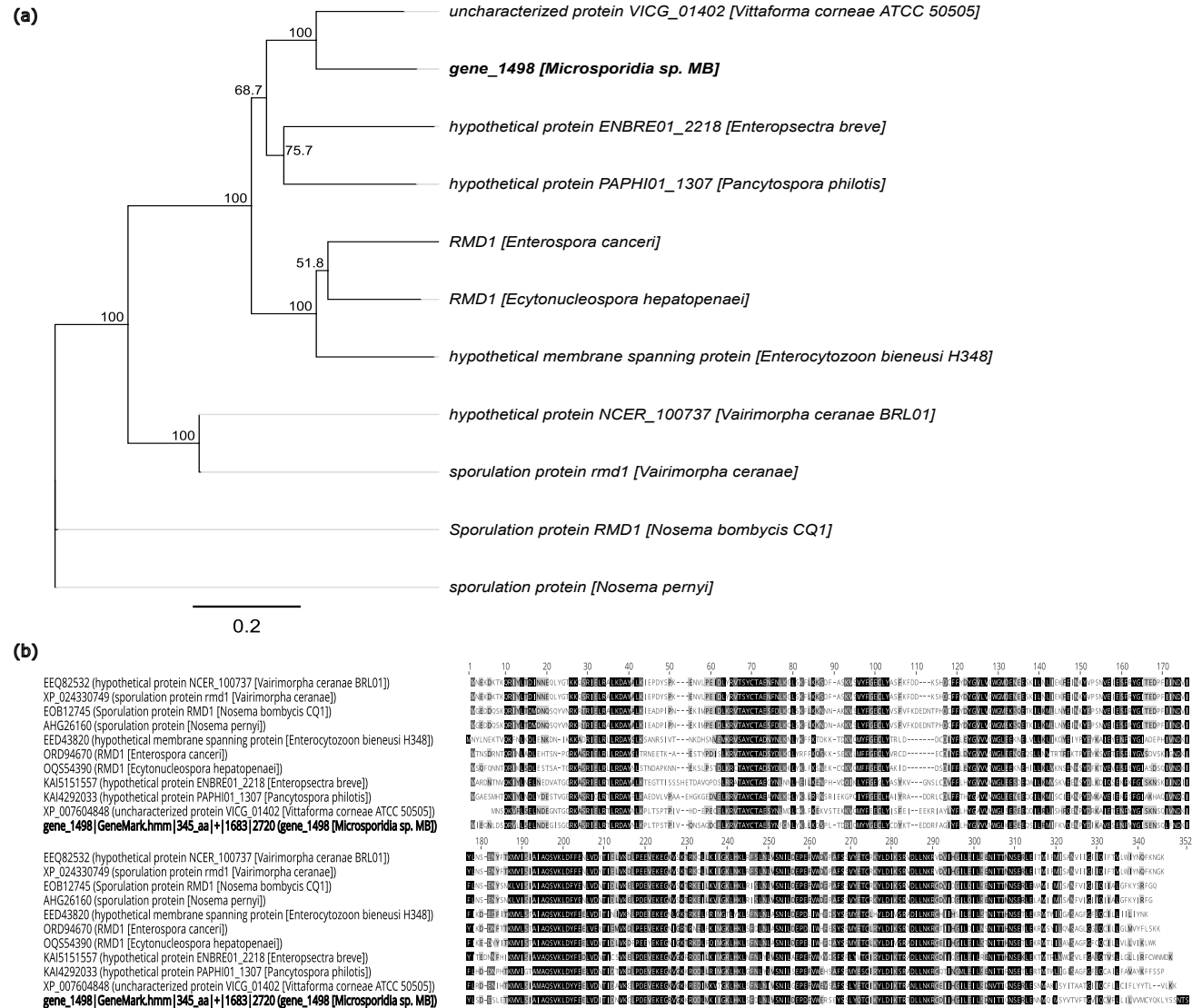


Figure 4.31: Tree calculations and sequence analysis of the Sporulation protein RMD1.

4.5.7 Phylogenetic Analysis of Iron-Sulfur Cluster Assembly Proteins in *Microsporidia* sp. MB

The Fe-S cluster assembly machinery in microsporidians involves vital proteins essential for Fe-S cluster biosynthesis and protein maturation. These proteins include Iron-Sulfur Cluster Scaffold Protein (Isu1), Cysteine (Nfs1), Frataxin (Yfh1), and heat shock protein 70 (HSP70). Isu1 serves as a scaffold protein for de novo synthesis of Fe-S clusters in yeast mitochondria. Nfs1 provides sulfur for Fe-S cluster biosynthesis as a putative iron donor in the Fe-S cluster assembly machinery. Mitochondrial HSP70 facilitates the transfer of Fe-S clusters from scaffold proteins to target apoproteins [[Huang et al., 2021](#), [Goldberg et al., 2008](#)].

These proteins play critical roles in the biosynthesis and assembly of Fe-S clusters, which are essential for the maturation of both mitochondrial and cytosolic proteins in microsporidians. Five orthologs linked to the Fe-S cluster assembly were identified in *Microsporidia* sp. MB (Table [4.2](#)). These were gene_19, identified as mitochondrial ferredoxin (Figure [4.32](#)); gene_371, homologous to cysteine desulfurase (Figure [4.33](#)), gene_1659 identified as the molecular chaperon HscB (Figure [4.34](#)); and gene_1773 as the iron-sulfur cluster assembly protein 1 (Figure [4.35](#)).

Table 4.2: An overview of *Microsporidia* sp. MB genes linked to the Fe-S cluster assembly machinery

Gene ID	Annotation GO Term	InterPro Name	Top BLAST Hit [species] (E-value; percent identity; query cover)
gene_19	mitochondrion;2 iron, 2 sulfur cluster binding;P450-containing electron transport chain	2Fe-2S ferredoxin-type iron-sulfur binding domain	Ferredoxin-2, mitochondrial [<i>Enteropsectra breve</i>] (2.38e-51;62%;98.37%)
gene_371	mitochondrion;cytosol;pyridoxal phosphate binding;cysteine desulfurase activity;[2Fe-2S] cluster assembly	Cysteine desulfurase	Cysteine desulfurase [<i>Pancycospora philotis</i>] (0; 71.1%; 100%)
gene_1659	ATPase activator activity;mitochondrion;[2Fe- 2S] cluster assembly;protein-folding chaperone binding	Chaperone J-domain superfamily	Molecular chaperone HscB [<i>Enteropsectra breve</i>] (3.53e-31; 41.2%; 90.53%)
gene_1773	cytoplasm;intracellular iron ion homeostasis;ferrous iron binding;iron-sulfur cluster assembly	Iron-sulfur cluster assembly scaffold protein IscU	iron-sulfur cluster assembly protein 1 [<i>V. corneae</i>] (3.67e-71; 79.4%; 97.76%)

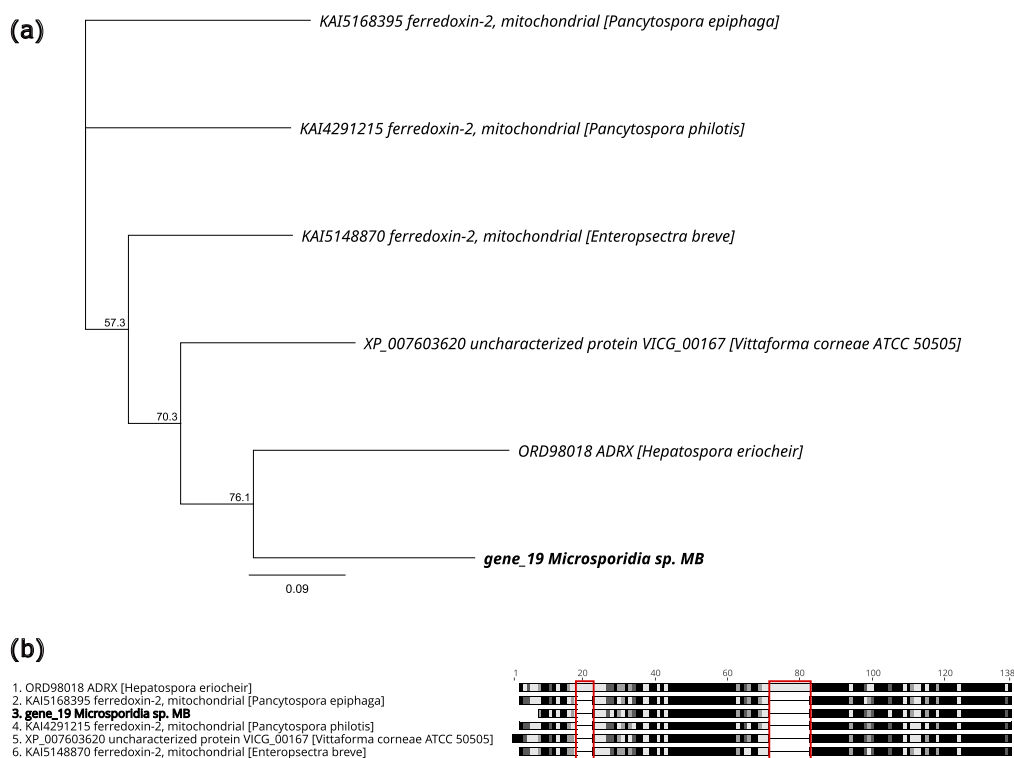


Figure 4.32: Analysis of the mitochondrial ferredoxin homolog in *Microsporidia* sp. MB gene_19. (a) Tree calculations done using Neighbor-joining method shows high similarity within this group. Bootstrap support values are highlighted on each branch, and the scale is shown at the bottom of the tree. (b) Multiple sequence alignment of gene_19 and its mitochondrial ferredoxin homologs. Two inserts were identified in *H. eriocheir*

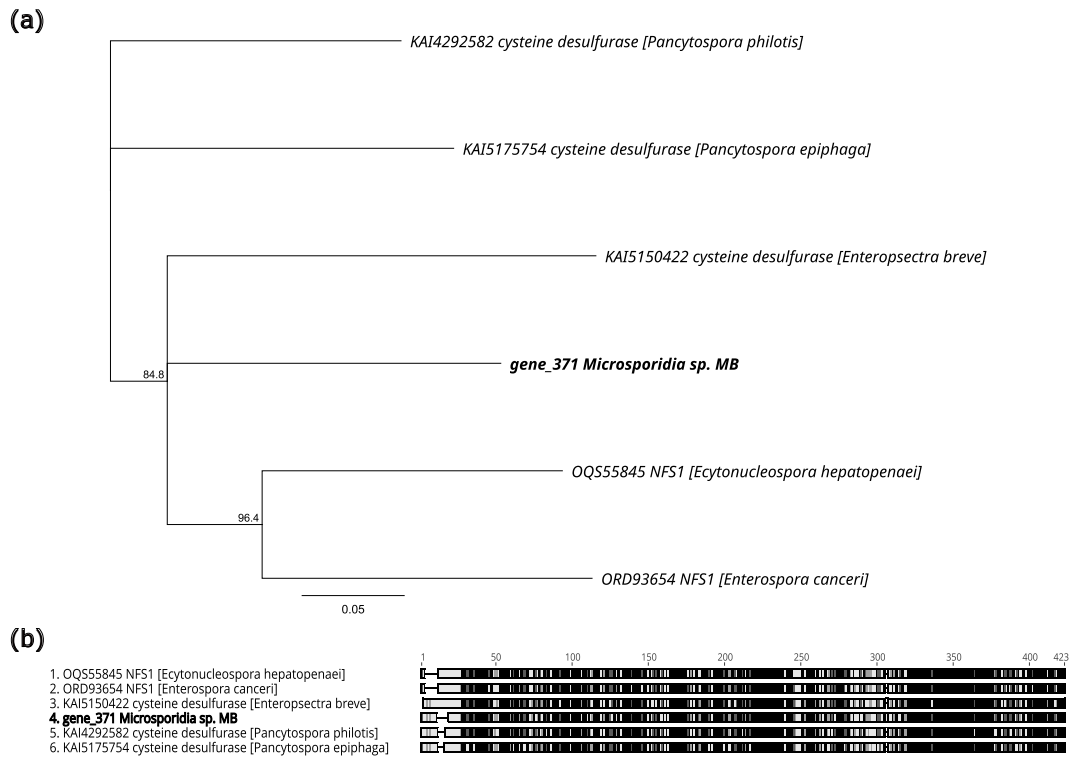


Figure 4.33: Analysis of the cysteine desulfurase homolog in *Microsporidia* sp. MB gene_371. (a) Tree calculations using the neighbour-joining method show high similarity. Bootstrap support values are highlighted on each branch, and the scale is shown at the bottom of the tree. (b) Multiple sequence alignment of gene_371 and its homologs show high similarity.

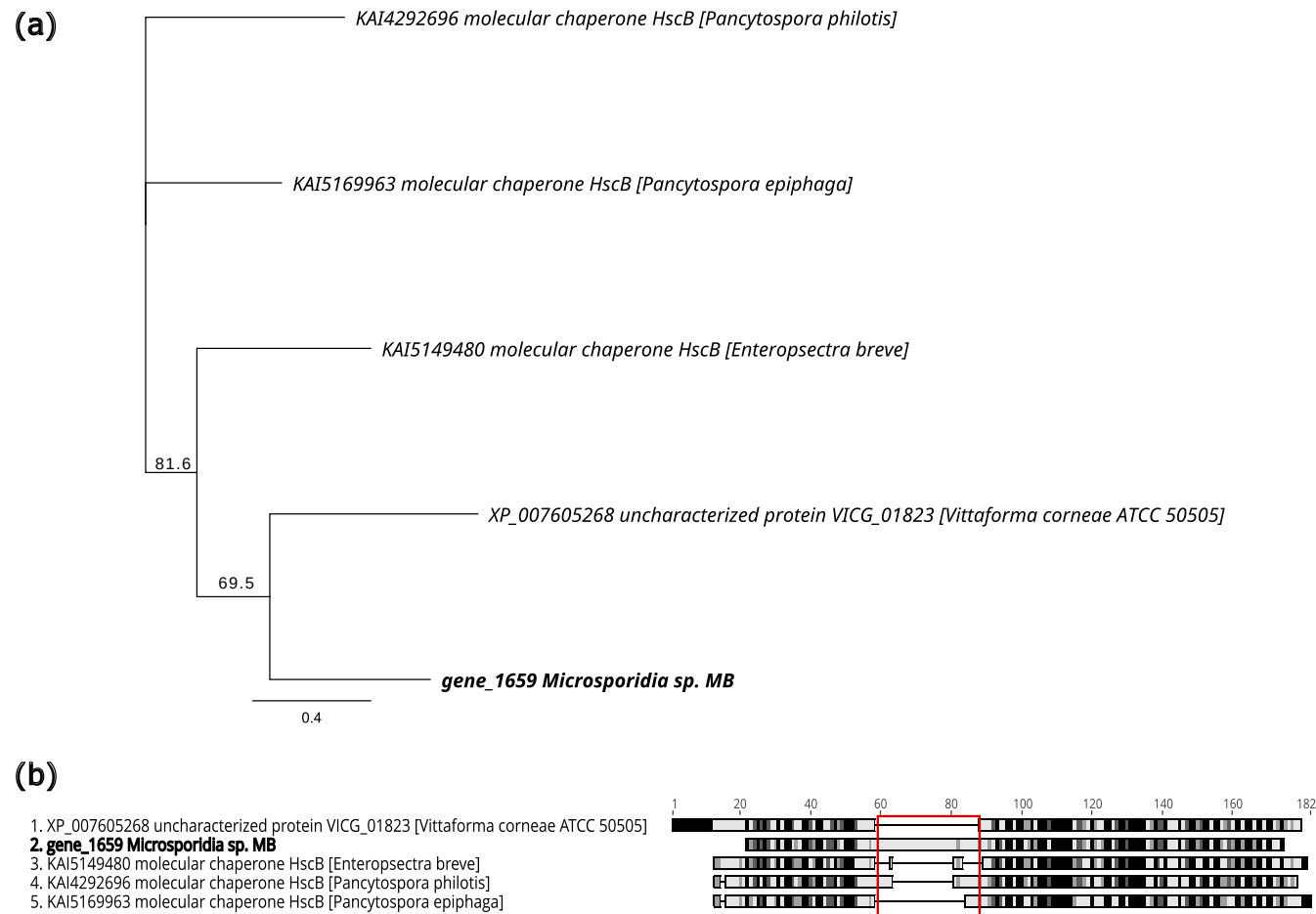


Figure 4.34: Analysis of the molecular chaperone HscB homolog in *Microsporidia* sp. MB gene_1659. (a) The generated neighbour-joining trees show high similarity. Bootstrap support values are indicated on each branch. The scale is shown at the bottom of the tree. (b) Multiple sequence alignment of gene_1659, highlighting an insert in *Microsporidia* sp. MB (red box).

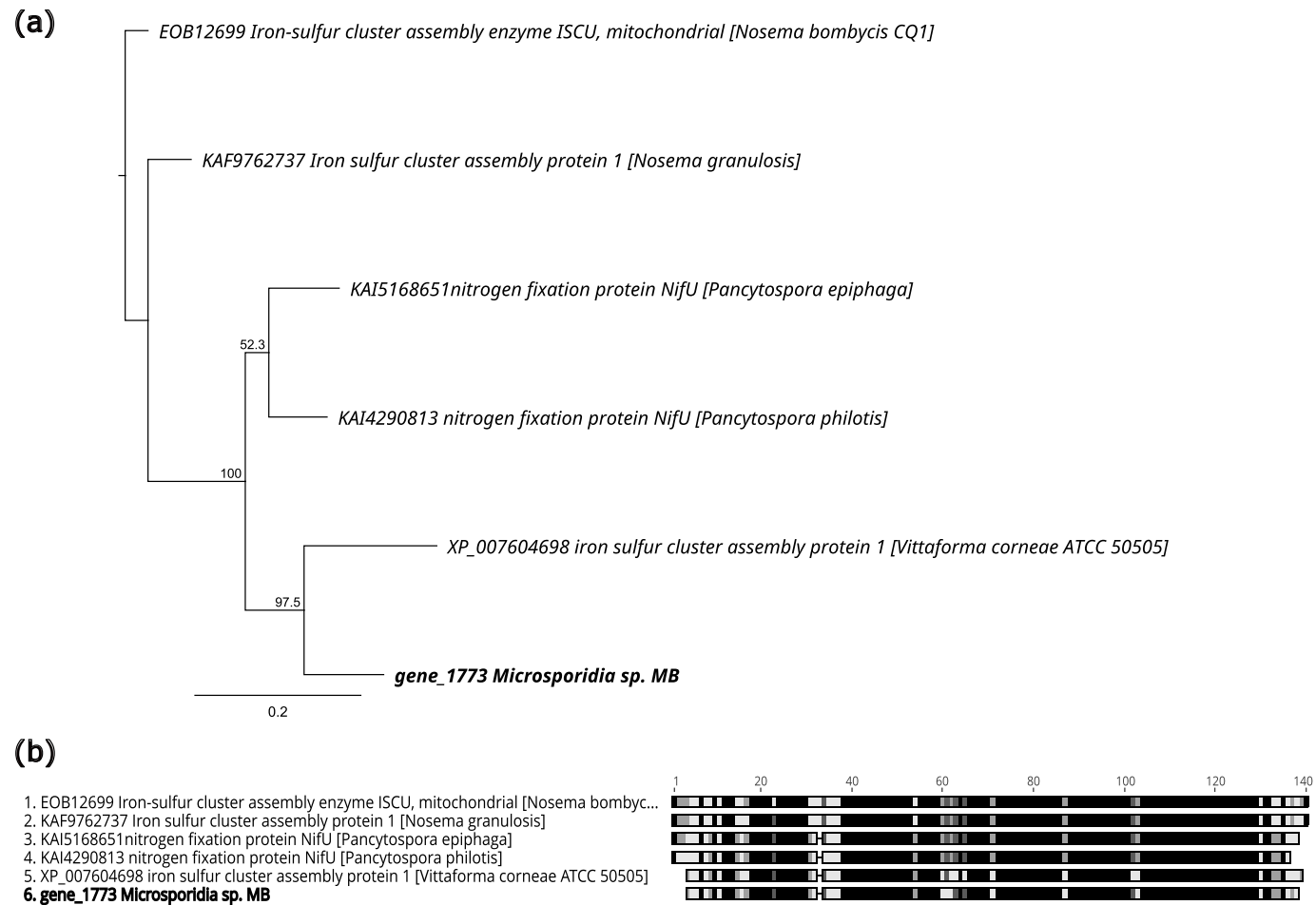


Figure 4.35: Analysis of the Fe-S cluster assembly enzyme IscU homolog in *Microsporidia* sp. MB gene_1773. (a) Tree calculations done with neighbour-joining method is illustrated here. Bootstrap support values are shown in each branch indicating high similarity. The scale is shown at the bottom of the tree. (b) Multiple sequence alignment of gene_1773 and other

4.6 Conclusion

This chapter comprehensively analyzes the genetic makeup of the *Microsporidia* sp. MB genome. It describes critical pathways in the symbiont and their gene conservation compared to other microsporidia species, providing a detailed insight into the genetic structure of this organism and its evolutionary conservation. The research emphasizes the significance of specific genes and pathways for the survival and proliferation of this microsporidia. This could potentially be useful for developing and applying this symbiont as a malaria biocontrol tool. The study revealed that *Microsporidia* sp. MB has retained all but one gene in the glycolytic pathway, suggesting its capacity to produce ATP independently. Multiple MRG orthologs, which are crucial components in meiosis, were identified in this microorganism. The presence of the RNAi machinery and the absence of many TEs highlight the role of RNAi in regulating the distribution of these repetitive elements in the genome. These findings altogether lay the foundation for future studies on host-symbiont interactions.

Chapter 5

Importance of Research, Future Directions, and Concluding Remarks

The pipeline developed for the re-annotation of the *V. corneae* genome provided a strong foundation for the annotation of the *Microsporidia* sp. MB genome. Since *V. corneae* and *Microsporidia* sp. MB share a close phylogenetic relationship, the insights gained from annotating *V. corneae*, particularly the identification and characterization of hypothetical proteins, were instrumental in guiding the annotation of *Microsporidia* sp. MB. The pipeline, which was designed to efficiently predict protein structure and function, allowed for the identification of conserved domains and functional motifs in *Microsporidia* sp. MB that might otherwise have been overlooked. Additionally, success of the pipeline with *V. corneae* provided a tested framework for handling complex genomic data, refining annotation accuracy, and focusing on key biological pathways, which were similarly explored in *Microsporidia* sp. MB. By applying the same rigorous computational approach, the annotation of *Microsporidia* sp. MB was both streamlined and enhanced, allowing for deeper exploration of its genome, particularly in relation to its unique biological processes such as meiosis and RNA interference. This continuity ensured that the re-annotation of *V. corneae* was not only a standalone project but directly beneficial for the genome assembly and annotation of *Microsporidia* sp. MB.

5.1 Chapter 2: Structural and Functional Annotation of Hypothetical Proteins from *Vittaforma corneae* ATCC 50505 Using in silico Approaches

5.1.1 Importance of Research

Microsporidia have lost an array of "unnecessary" genes through evolution and retained only a small subset of genes they require to survive in their host systems. Most of the proteins found in these intracellular parasites are often left undescribed, having no known function. However, these proteins have been suggested to be pivotal in the survival and proliferation of these organisms. The availability of several sequenced genomes of these complex microorganisms. Accurate and proper annotation of HPs is important to translate information from sequences to understand new functions, interactions, structures, and pathways in which these unknown proteins might be involved. Developing a systemic computational workflow to annotate these proteins is helpful to inform wet lab experiments that are often expensive, time-consuming, and laborious.

5.1.2 Summary of Novel Findings

Predicting HP function, conserved domains, involvement in several biological pathways, and establishing evolutionary relationships between different proteins by employing a systematic and rigorous computational annotation pipeline provides a comprehensive overview of *V. corneae* biological system. As with most microsporidia, a large portion of the proteins in this organism were involved in regulatory and transport functions, critical for the survival of microsporidia due to their compact genomes and extreme gene loss. Several conserved functional domains are highlighted in this parasite and its relatives in the same clade of microsporidia and could be used in the taxonomic classification of microsporidia. An extensive analysis of PPI networks, phylogeny, and motif calculations, coupled

with homology modeling of several case studies, demonstrates the capacity of these in silico methods to describe the structure and function of unknown proteins.

5.1.3 Further Implications and Future Directions

The annotation pipeline described in this section can be extrapolated into characterizing proteins from newly sequenced genomes. Online databases and resources are constantly being updated and curated. Providing a simple yet precise pipeline to annotate proteins using these tools will enable a better understanding of the studied biological system, where limited lab resources are required. These findings could have significant implications for drug discovery and other biomedical applications. Moreover, recent breakthroughs in computational biology, particularly in 3D protein modeling, are significantly advancing our understanding of microorganisms like Microsporidia. Tools such as AlphaFold [Ronneberger et al., 2021], with its remarkable accuracy in predicting protein structures, are enabling researchers to explore the functional roles of proteins in these parasitic organisms in ways that were previously unattainable. Complementary structural search tools like FoldSeek [van Kempen et al., 2022] are becoming indispensable for comparing vast numbers of protein structures, helping to uncover evolutionary relationships and molecular mechanisms at unprecedented scales.

5.1.4 Concluding Remarks

Multiple computational tools and online resources are readily available and provide a more cost-effective and efficient approach to annotating biomolecules before conducting laboratory experiments. The findings in this chapter highlight an efficient step-by-step annotation framework to understand protein structure and function. This method can significantly improve our understanding of complex biological systems at a lower cost.

5.2 Chapter 3: De novo Genome Assembly of *Microsporidia* sp. MB isolated from *Anopheles arabiensis* in Kenya

5.2.1 Importance of Research

The study of microsporidia genomes has provided valuable insights into these organisms' biology, evolution, and pathogenicity. Evidence of symbiotic microsporidia in mosquitoes that can block malaria transmission offers promising opportunities in the global fight against malaria. Isolating, assembling, and describing the genome structure of this symbiont sets up research aimed at understanding its biological mechanism in conferring this protection in its host.

5.2.2 Summary of Novel Findings

The *Microsporidia* sp. MB genome has a size of 5.9Mb and has a small proportion of repetitive regions. Transposable elements have unique functions in microsporidia, including stabilizing the genome structure, host specificity, genomic variation across different species, and regulation of several genes. The absence of long repeats in this genome explains its relatively small size. Identifying several short tandem repeats enables further studies into this organism's genetic variation and evolutionary studies across different host species. The classification of microsporidia species is complex and continuously shifting, often relying on the SSU rRNA gene and ultrastructural characterization. Comparative phylogenomic analysis groups *Microsporidia* sp. MB close to *V. corneae* in the Enterocytozoonida clade. Using a multi-gene approach in classifying this organism further strengthens previous reports on its taxonomy. This underscores the potential to identify more molecular markers that could be used in the comprehensive classification of microsporidia.

5.2.3 Further Implications and Future Directions

The availability of the reference genome of *Microsporidia* sp. MB, fast development of cost-effective sequencing technologies, and improvements in bioinformatics approaches will make next-generation sequencing of this symbiont easier and unravel the ever-rising questions on phylogeny and genetic diversity of microsporidia species. Advancements in sequencing technologies are on the brink of revolutionizing Microsporidia genomics. Enhanced long-read sequencing platforms, coupled with more sophisticated genome assembly algorithms, are expected to soon deliver chromosomal-level assemblies of Microsporidia genomes, including the elusive *Microsporidia* sp. MB. These high-resolution assemblies will provide critical insights into the organism's complex biology, host-parasite interactions, and potential targets for therapeutic intervention, marking a new era in parasitic disease research.

5.2.4 Concluding Remarks

The analysis of the genome structure and content of *Microsporidia* sp. MB forms the basis for acquiring a better understanding of the biology of infection as a first step into its application as a biocontrol tool. In this regard, we isolated, sequenced, and assembled the first draft genome of *Microsporidia* sp. MB and demonstrated its similarity to other microsporidia in the same clade. Incomplete genome sequencing or partial genome assembly could result in missed gene annotations due to sequence gaps. While this study aimed for a comprehensive assembly, there remains a possibility that important genes, including those associated with LRR motifs, were not fully captured. Future work should focus on confirming the completeness of the genome assembly, including a hybrid assembly using long-read sequencing platforms and addressing any potential gaps to ensure accurate gene annotations.

5.3 Chapter 4: *Microsporidia* sp. MB Genome Annotation and Functional Analysis

5.3.1 Importance of Research

Understanding the physiological and biological characteristics of *Microsporidia* sp. MB and its intricate relationship with the *Anopheles* mosquito are crucial for understanding its role in malaria control. Annotating the genome of this symbiont provides important information on its genetic makeup and highlights evidence of unique gene features and biochemical processes in the symbiont.

5.3.2 Summary of Novel Findings

The investigation of the gene content of *Microsporidia* sp. MB genome highlighted that this symbiont is similar to *V. corneae* and includes a range of unique features. This genome contains 2,247 putative genes, and their exact functional annotation is based solely on sequence homology, forming a baseline for in vivo and in vitro molecular experiments. A large frequency of proteins involved in enzymatic activities was described. Unlike most genomes in Enterocytozoonida microsporidia, this symbiont retained all but one gene involved in the glycolytic pathway, similar to its sister *V. corneae*. The presence of the three main features of the RNAi machinery and several MRGs suggest host interaction and mode of sexual reproduction. This research highlights the biology of *Microsporidia* sp. MB and fundamentally provides valuable resources to further enable studies into the molecular mechanisms of host-symbiont interactions.

5.3.3 Further Implications and Future Directions

The study of the *Microsporidia* sp. MB genome has provided a foundation for future research in evolutionary biology, symbiosis, and host-microbe interactions. The variations in gene content among microsporidia species emphasize the need for further research to understand the underlying mechanisms driving this diversity and its impact on the adaptive strategies of these organisms. The annotation of

the protein content. RNASeq and other NGS analyses could further be used to validate these annotations. Transcriptomics analysis at different life stages of the symbiont and samples isolated in different tissues could provide crucial information on the potential roles of these genes. This approach could also highlight genes likely involved in the interactions of tripartite host-symbiont-parasite (*Anopheles-Microsporidia-Plasmodium*).

5.3.4 Concluding Remarks

The analysis of the genome of *Microsporidia* sp. MB establishes the foundation for multiple scientific studies. Similar to observations in closely related microsporidia, *Microsporidia* sp. MB has retained several genes in the glycolytic pathway, MRGs, and main RNAi components. The findings observed here form the basis for understanding the basic biology of the symbiont and host-symbiont interactions.

List of References

- Jewelna Akorli, Esinam Abba Akorli, Seraphim Naa Afoley Tetteh, Godwin Kwame Amlalo, Millicent Opoku, Rebecca Pwalia, Michelle Adimazoya, Dorcas Atibilla, Sellase Pi-Bansa, Joseph Chabi, and Samuel Kweku Dadzie. Microsporidia MB is found predominantly associated with *Anopheles gambiae* s.s and *Anopheles coluzzii* in Ghana. *Scientific Reports*, 11(1):1–5, 2021. ISSN 20452322. doi: 10.1038/s41598-021-98268-2. URL <https://doi.org/10.1038/s41598-021-98268-2>.
- William G. Alexander, Jennifer H. Wisecaver, Antonis Rokas, and Chris Todd Hittinger. Horizontally acquired genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, 113(15):4116–4121, 2016. ISSN 10916490. doi: 10.1073/pnas.1517242113.
- Matthew T. Aliota, Emma C. Walker, Alexander Uribe Yepes, Ivan Dario Velez, Bruce M. Christensen, and Jorge E. Osorio. The wMel Strain of *Wolbachia* Reduces Transmission of Chikungunya Virus in *Aedes aegypti*. *PLoS Neglected Tropical Diseases*, 2016. ISSN 19352735. doi: 10.1371/journal.pntd.0004677.
- José Juan Almagro Armenteros, Konstantinos D. Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, 37(4):420–423, 2019. ISSN 15461696. doi: 10.1038/s41587-019-0036-z.
- Theodore G. Andreadis. *Amblyospora connecticus* sp. nov. (Microsporida: Amblyosporidae): Horizontal transmission studies in the mosquito *Aedes cantator*

- and formal description. *Journal of Invertebrate Pathology*, 52(1):90–101, 1988. ISSN 10960805. doi: 10.1016/0022-2011(88)90107-3.
- Theodore G. Andreadis. Epizootiology of *Amblyospora connecticus* (Microsporida) in Field Populations of the Saltmarsh Mosquito, *Aedes cantator*, and the Cyclopoid Copepod, *Acanthocyclops vernalis*. *The Journal of Protozoology*, 37(3):174–182, 1990. ISSN 15507408. doi: 10.1111/j.1550-7408.1990.tb01123.x.
- Theodore G. Andreadis. Microsporidian Parasites of Mosquitoes. *Journal of the American Mosquito Control Association*, 23(sp2):3–29, 2007. ISSN 8756-971X. doi: 10.2987/8756-971X(2007)23[3:MPOM]2.0.CO;2.
- Simon Andrews. FastQC a Quality Control Tool for High Throughput Sequence Data, 2010. URL <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Lilian Mbaisi Ang’ang’o, Jeremy Keith Herren, and Özlem Tastan Bishop. Structural and Functional Annotation of Hypothetical Proteins 2 from the Microsporidia Species *Vittaforma corneae* ATCC 50505 3 Using in silico Approaches. *Int. J. Mol. Sci*, 2022, 2022. doi: 10.3390/xxxxx.
- Lilian Mbaisi Ang’ang’o, Jacqueline Wahura Waweru, Edward Edmond Makhulu, Anne Wairimu, Fidel Gabriel Otieno, Thomas Onchuru, Özlem Tastan Bishop, and Jeremy Keith Herren. Draft genome of microsporidia sp. mb—a malaria-blocking microsporidian symbiont of the anopheles arabiensis. *Microbiology Resource Announcements*, 13(4):e00903–23, 2024.
- Lilian Mbaisi Ang’ang’o, Jeremy Keith Herren, and Özlem Tastan Bishop. Structural and Functional Annotation of Hypothetical Proteins from the Microsporidia Species *Vittaforma corneae* ATCC 50505 Using in silico Approaches. *International Journal of Molecular Sciences*, 24(4), 2 2023. ISSN 14220067. doi: 10.3390/ijms24043507.
- Thomas H. Ant, Christie S. Herd, Vincent Geoghegan, Ary A. Hoffmann, and Steven P. Sinkins. The Wolbachia strain wAu provides highly efficient virus transmission blocking in *Aedes aegypti*. *PLoS Pathogens*, 14(1):1–19, 2018. ISSN 15537374. doi: 10.1371/journal.ppat.1006815.

- Jose Juan Almagro Armenteros, Marco Salvatore, Olof Emanuelsson, Ole Winther, Gunnar Von Heijne, Arne Elofsson, and Henrik Nielsen. Detecting sequence signals in targeting peptides using deep learning. *Life science alliance*, 2(5), 2019.
- Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, and Gavin Rubin Gerald M. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 5 2000. ISSN 1061-4036. doi: 10.1038/75556. URL https://www.nature.com/articles/ng0500_25.
- Terri K. Attwood. The PRINTS database: a resource for identification of protein families. *Briefings in bioinformatics*, 3(3):252–263, 2002. ISSN 14675463. doi: 10.1093/bib/3.3.252.
- Cristina Aurrecochea, Ana Barreto, John Brestelli, Brian P. Brunk, Elisabet V. Caler, Steve Fischer, Bindu Gajria, Xin Gao, Alan Gingle, Greg Grant, Omar S. Harb, Mark Heiges, John Iodice, Jessica C. Kissinger, Eileen T. Kraemer, Wei Li, Vishal Nayak, Cary Pennington, Deborah F. Pinney, Brian Pitts, David S. Roos, Ganesh Srinivasamoorthy, Christian J. Stoeckert, Charles Treatman, and Haiming Wang. AmoebaDB and MicrosporidiaDB: Functional genomic resources for Amoebozoa and Microsporidia species. *Nucleic Acids Research*, 39 (SUPPL. 1):612–619, 2011. ISSN 03051048. doi: 10.1093/nar/gkq1006.
- Susan W Avery and Albert H Undeen. Horizontal Transmission of Parathelohania anophelis to the Anopheles quadrimaculatus”. *Journal of Invertebrate Pathology*, 56:98–105, 1990.
- Timothy L Bailey. Discovering sequence motifs. *Comparative Genomics*, pages 271–292, 2008.
- Timothy L. Bailey and Michael Gribskov. Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics*, 14(1):48–54, 1998. ISSN 13674803. doi: 10.1093/bioinformatics/14.1.48.

- Timothy L. Bailey, Nadya Williams, Chris Misleh, and Wilfred W. Li. MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34(WEB. SERV. ISS.):369–373, 2006. ISSN 03051048. doi: 10.1093/nar/gkl198.
- Timothy L. Bailey, James Johnson, Charles E. Grant, and William S. Noble. The MEME Suite. *Nucleic Acids Research*, 43(W1):W39–W49, 2015. ISSN 13624962. doi: 10.1093/nar/gkv416.
- Bachir Balech, Alfonso Monaco, Michele Perniola, Monica Santamaria, Giacinto Donvito, Saverio Vicario, Giorgio Maggi, and Graziano Pesole. DNA Multiple Sequence Alignment Guided by Protein Domains: The MSA-PAD 2.0 Method. *Viral Metagenomics: Methods and Protocols*, pages 173–180, 2018.
- Alexander R Ball Jr and Kyoko Yokomori. The structural maintenance of chromosomes (SMC) family of proteins in mammals. *Chromosome Research*, 9(2): 85–96, 2001.
- Hironori Bando, Kiyoshi Okado, Wamdaogo M Guelbeogo, Athanase Badolo, Hiroka Aonuma, Bryce Nelson, Shinya Fukumoto, Xuenan Xuan, N’fale Sagnon, and Hirotaka Kanuka. Intra-specific diversity of *Serratia marcescens* in *Anopheles* mosquito midgut defines *Plasmodium* transmission capacity. *Scientific reports*, 3, 2013. ISSN 2045-2322. doi: 10.1038/srep01641.
- Irka Bargielowski and Jacob C. Koella. A possible mechanism for the suppression of *Plasmodium berghei* development in the mosquito *Anopheles gambiae* by the Microsporidian *Vavraia culicis*. *PLoS ONE*, 4(3), 2009. ISSN 19326203. doi: 10.1371/journal.pone.0004676.
- Alex Bateman. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1), 2019. ISSN 13624962. doi: 10.1093/nar/gky1049.
- Kelly S. Bateman, Rose Kerr, Grant D. Stentiford, Tim P. Bean, Chantelle Hooper, Benigna Van Eynde, Daan Delbare, Jamie Bojko, Olivier Christiaens, Clauvis N.T. Taning, Guy Smagghe, Monique M. van Oers, and Ronny van

- Aerle. Identification and full characterisation of two novel crustacean infecting members of the family nudiviridae provides support for two subfamilies. *Viruses*, 13(9):1–25, 2021. ISSN 19994915. doi: 10.3390/v13091694.
- J J Becnel and T G Andreadis. Microsporidia in Insects. In *The Microsporidia and Microsporidiosis*, pages 447–501. 1999.
- James J. Becnel, Susan E. White, and Alexandra M. Shapiro. Review of microsporidia-mosquito relationships: From the simple to the complex. *Folia Parasitologica*, 52(1-2):41–50, 2005. ISSN 00155683. doi: 10.14411/fp.2005.006.
- Pascal Benkert, Michael Künzli, and Torsten Schwede. QMEAN server for protein model quality estimation. *Nucleic Acids Research*, 37(SUPPL. 2), 2009. ISSN 03051048. doi: 10.1093/nar/gkp322.
- Marcio V Bertacine Dias, Jademilson C Santos, Gerardo A Libreros-Zuniga, Joao A Ribeiro, and Sair M Chavez-Pacheco. Folate biosynthesis pathway: mechanisms and insights into drug design for infectious diseases. *Future medicinal chemistry*, 10(8):935–959, 2018.
- John Besemer, Alexandre Lomsadze, and Mark Borodovsky. GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, 29(12):2607–2618, 2001. ISSN 03051048. doi: 10.1093/nar/29.12.2607.
- Jeremy S Bickel, Liting Chen, Jin Hayward, Szu Ling Yeap, Ashley E Alkers, and Raymond C Chan. Structural maintenance of chromosomes (SMC) proteins promote homolog-independent recombination repair in meiosis crucial for germ cell genomic stability. *PLoS genetics*, 6(7):e1001028, 2010.
- Susan Bjørnson and David Oi. Microsporidia Biological Control Agents and Pathogens of Beneficial Insects. *Microsporidia: Pathogens of Opportunity: First Edition*, pages 635–670, 2014. doi: 10.1002/9781118395264.ch25.
- Matthias Blum, Hsin Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasaamy, Alex Mitchell, Gift Nuka, Typhaine Paysan-Lafosse, Matloob Qureshi, Shriya Raj, Lorna Richardson, Gustavo A. Salazar, Lowri Williams,

- Peer Bork, Alan Bridge, Julian Gough, Daniel H. Haft, Ivica Letunic, Aron Marchler-Bauer, Huaiyu Mi, Darren A. Natale, Marco Necci, Christine A. Orengo, Arun P. Pandurangan, Catherine Rivoire, Christian J.A. Sigrist, Ian Sillitoe, Narmada Thanki, Paul D. Thomas, Silvio C.E. Tosatto, Cathy H. Wu, Alex Bateman, and Robert D. Finn. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, 49(D1):D344–D354, 2021. ISSN 13624962. doi: 10.1093/nar/gkaa977.
- Ulrich Bodenhofer, Enrico Bonatesta, Christoph Horejš-Kainrath, and Sepp Hochreiter. msa: an R package for multiple sequence alignment. *Bioinformatics*, 31(24):3997–3999, 2015.
- Anne Boissière, Majoline T. Tchioffo, Dipankar Bachar, Luc Abate, Alexandra Marie, Sandrine E. Nsango, Hamid R. Shahbazkia, Parfait H. Awono-Ambene, Elena A. Levashina, Richard Christen, and Isabelle Morlais. Midgut microbiota of the malaria mosquito vector *Anopheles gambiae* and interactions with *Plasmodium falciparum* infection. *PLoS Pathogens*, 8(5):1–12, 2012. ISSN 15537366. doi: 10.1371/journal.ppat.1002742.
- Jamie Bojko and Mykola Ovcharenko. Pathogens and other symbionts of the Amphipoda: Taxonomic diversity and pathological significance. *Diseases of Aquatic Organisms*, 136(1):3–36, 2019. ISSN 16161580. doi: 10.3354/dao03321.
- Jamie Bojko, Aaron W. Reinke, Grant D. Stentiford, Bryony Williams, Martin S.J. Rogers, and David Bass. Microsporidia: a new taxonomic, evolutionary, and ecological synthesis. *Trends in Parasitology*, 38(8):642–659, 2022. ISSN 14715007. doi: 10.1016/j.pt.2022.05.007. URL <https://doi.org/10.1016/j.pt.2022.05.007>.
- Pablo Bolaños-Villegas. The role of structural maintenance of chromosomes complexes in meiosis and genome maintenance: Translating biomedical and model plant research into crop breeding opportunities. *Frontiers in Plant Science*, 12: 659558, 2021.
- Lakhyajit Boruah, Aparoop Das, Lalit Mohan Nainwal, Neha Agarwal, and Brajesh Shankar. In-Silico Drug Design: A revolutionary approach to change the

- concept of current Drug Discovery Process. *Indian Journal of Pharmaceutical and Biological Research*, 1(02):60, 2013. doi: 10.30750/ijpbr.1.2.11.
- Broad Institute. Vittaforma corneae ATCC 50505 Genome sequencing and assembly., 2011. URL <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA63501>.
- David K Brown. Bioinformatics tool development with a focus on structural bioinformatics and the analysis of genetic variation in humans by. (September), 2017.
- Tullu Bukhari, Roland Pevsner, and Jeremy Keith. Herren. Microsporidia: a promising vector control tool for residual malaria transmission. *Frontiers in Tropical Diseases*, 3(July):1–18, 2022. doi: 10.3389/fitd.2022.957109.
- Judy F. Burnham. Scopus database: A review. *Biomedical Digital Libraries*, 3: 1–8, 2006. ISSN 17425581. doi: 10.1186/1742-5581-3-1.
- Lena Burri, Bryony A P Williams, Dejan Bursac, Trevor Lithgow, and Patrick J Keeling. Microsporidian mitosomes retain elements of the general mitochondrial targeting system. 2006. doi: 10.1073/pnas.0604109103.
- Ann Cali and Peter M. Takvorian. Developmental Morphology and Life Cycles of the Microsporidia. *Microsporidia: Pathogens of Opportunity: First Edition*, pages 71–133, 2014. doi: 10.1002/9781118395264.ch2.
- Scott E. Campbell, Tom A. Williams, Asim Yousuf, Darren M. Soanes, Konrad H. Paszkiewicz, and Bryony A.P. Williams. The Genome of *Spraguea lophii* and the Basis of Host-Microsporidian Interactions. *PLoS Genetics*, 9(8):1–15, 2013. ISSN 15537390. doi: 10.1371/journal.pgen.1003676.
- Cintia Cansado-Utrilla, Serena Y. Zhao, Philip J. McCall, Kerri L. Coon, and Grant L. Hughes. The microbiome and mosquito vectorial capacity: rich potential for discovery and translation. *Microbiome*, 9(1):1–11, 2021. ISSN 20492618. doi: 10.1186/s40168-021-01073-2.
- Salvador Capella-Gutiérrez, José M Silla-Martínez, and Toni Gabaldón. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973, 2009.

- Salvador Capella-Gutiérrez, Marina Marcet-Houben, and Toni Gabaldón. Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. *BMC Biology*, 10(1):47, 2012. ISSN 1741-7007. doi: 10.1186/1741-7007-10-47. URL <http://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-10-47>.
- Eric P. Caragata and Thomas Walker. Using bacteria to treat diseases. *Expert Opinion on Biological Therapy*, 12(6):701–712, 2012. ISSN 1471-2598. doi: 10.1517/14712598.2012.677429. URL <http://www.tandfonline.com/doi/full/10.1517/14712598.2012.677429>.
- Jeffrey T Chang, Soumya Raychaudhuri, and Russ B Altman. Including biological literature improves homology search. In *Biocomputing 2001*, pages 374–383. World Scientific, 2000.
- T Mfangnia Charlene N, Henri E Z Tonnang, Tsanou Berge, and Herren Jeremy. Mathematical modelling of the interactive dynamics of wild and Microsporidia MB-infected mosquitoes. 2023.
- Yuxin Chen, Yongsheng Chen, Chunmei Shi, Zhibo Huang, Yong Zhang, Shengkang Li, Yan Li, Jia Ye, Chang Yu, Zhuo Li, Xiuqing Zhang, Jian Wang, Huanming Yang, Lin Fang, and Qiang Chen. SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience*, 7(1):1–6, 2018. ISSN 2047217X. doi: 10.1093/gigascience/gix120.
- Mary J Clancy. Meiosis: step-by-step through sporulation. *Current Biology*, 8(13):R461–R463, 1998.
- Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, 2005.
- Nicolas Corradi. Microsporidia: Eukaryotic Intracellular Parasites Shaped by Gene Loss and Horizontal Gene Transfers. *Annual Review of Microbiology*, 69(1):167–183, 2015. ISSN 15453251. doi: 10.1146/annurev-micro-091014-104136.

- Nicolas Corradi and Mohammed Selman. Latest progress in microsporidian genome research. *Journal of Eukaryotic Microbiology*, 60(3):309–312, 2013. ISSN 10665234. doi: 10.1111/jeu.12030.
- Nicolas Corradi and Claudio H. Slamovits. The intriguing nature of microsporidian genomes. *Briefings in Functional Genomics*, 10(3):115–124, 2011. ISSN 20412649. doi: 10.1093/bfgp/elq032.
- Nicolas Corradi, Jean François Pombert, Laurent Farinelli, Elizabeth S. Didier, and Patrick J. Keeling. The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*. *Nature Communications*, 1(6), 2010. ISSN 20411723. doi: 10.1038/ncomms1082.
- Gualberto Asencio Cortés and Jesús A Aguilar-Ruiz. Predicting protein distance maps according to physicochemical properties. *Journal of Integrative Bioinformatics*, 8(3):158–175, 2011.
- Christina A. Cuomo, Christopher A. Desjardins, Malina A. Bakowski, Jonathan Goldberg, Amy T. Ma, James J. Becnel, Elizabeth S. Didier, Lin Fan, David I. Heiman, Joshua Z. Levin, Sarah Young, Qiandong Zeng, and Emily R. Troemel. Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth. *Genome Research*, 22(12):2478–2488, 2012. ISSN 10889051. doi: 10.1101/gr.142802.112.
- Nsa Dada, Natapong Jupatanakul, Guillaume Minard, Sarah M. Short, Jewelna Akorli, and Luis Martinez Villegas. Considerations for mosquito microbiome research from the Mosquito Microbiome Consortium. *Microbiome*, 9(1):1–16, 2021. ISSN 20492618. doi: 10.1186/s40168-020-00987-7.
- Petr Danecek, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, Thomas Keane, Shane A. McCarthy, and Robert M. Davies. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):1–4, 2021. ISSN 2047217X. doi: 10.1093/gigascience/giab008.
- Edouard de Castro, Christian J.A. Sigrist, Alexandre Gattiker, Virginie Buliard, Petra S. Langendijk-Genevaux, Elisabeth Gasteiger, Amos Bairoch, and

- Nicolas Hulo. ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research*, 34(WEB. SERV. ISS.):362–365, 2006. ISSN 03051048. doi: 10.1093/nar/gkl124.
- Jesús De La Cruz, Dieter Kressler, Patrick Linder, J De La Cruz, D Kressler, and P Linder. Unwinding RNA in *Saccharomyces cerevisiae*: DEAD-box proteins and related families. 0004(99):192–198, 1999.
- Paul Dean, Robert P. Hirt, and T. Martin Embley. Microsporidia: Why Make Nucleotides if You Can Steal Them? *PLoS Pathogens*, 12(11):1–13, 2016. ISSN 15537374. doi: 10.1371/journal.ppat.1005870.
- F. Delbac, I. Peuvél, G. Metenier, E. Peyretailade, and C. P. Vivares. Microsporidian invasion apparatus: Identification of a novel polar tube protein and evidence for clustering of ptp1 and ptp2 genes in three *Encephalitozoon* species. *Infection and Immunity*, 69(2):1016–1024, 2001. ISSN 00199567. doi: 10.1128/IAI.69.2.1016-1024.2001.
- Teegan A. Dellibovi-Ragheb, Jolyn E. Gisselberg, and Sean T. Prigge. Parasites FeS Up: Iron-Sulfur Cluster Biogenesis in Eukaryotic Pathogens. *PLoS Pathogens*, 9(4), 4 2013. ISSN 15537366. doi: 10.1371/journal.ppat.1003227.
- Nathan J. Dennison, Natapong Jupatanakul, and George Dimopoulos. The mosquito microbiota influences vector competence for human pathogens. *Current Opinion in Insect Science*, 2014. ISSN 22145745. doi: 10.1016/j.cois.2014.07.004.
- Peter Deplazes, A Mathis, M van Saanen, Anne Iten, Robert Keller, Isabelle Tanner, Michel P Glauser, Rainer Weber, and Elizabeth U Canning. Dual microsporidial infection due to *Vittaforma corneae* and *Encephalitozoon hellem* in a patient with AIDS. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 27 6:1521–4, 1998. URL <https://api.semanticscholar.org/CorpusID:15371541>.
- Christopher a Desjardins, Neil D Sanscrainte, Jonathan M Goldberg, David

- Heiman, Sarah Young, Qiandong Zeng, Hiten D Madhani, James J Becnel, and Christina a Cuomo. Contrasting host-pathogen interactions and genome evolution in two generalist and specialist microsporidian pathogens of mosquitoes. *Nature communications*, 6(May):7121, 2015. ISSN 2041-1723. doi: 10.1038/ncomms8121. URL <http://www.ncbi.nlm.nih.gov/pubmed/25968466>.
- Claus Desler, Prashanth Suravajhala, May Sanderhoff, Merete Rasmussen, and Lene J. Rasmussen. In Silico screening for functional candidates amongst hypothetical proteins. *BMC Bioinformatics*, 10:289, 2009. ISSN 14712105. doi: 10.1186/1471-2105-10-289.
- Elizabeth S. Didier. Microsporidiosis: An emerging and opportunistic infection in humans and animals. *Acta Tropica*, 94(1):61–76, 2005. ISSN 0001706X. doi: 10.1016/j.actatropica.2005.01.010.
- Oswald Y. Djihinto, Adandé A. Medjigbodo, Albert R.A. A. Gangbadja, Helga M. Saizonou, Hamirath O. Lagnika, Dyane Nanmede, Laurette Djossou, Roméo Bohounton, Pierre Marie Sovegnon, Marie-Joel Joel Fanou, Romuald Agonhosou, Romaric Akoton, Wassiyath Mousse, and Luc S. Djogbénou. Malaria-Transmitting Vectors Microbiota: Overview and Interactions With Anopheles Mosquito Biology. *Frontiers in Microbiology*, 13(May), 2022. ISSN 1664302X. doi: 10.3389/fmicb.2022.891573.
- Mutlu Doğruel, Thomas A Down, and Tim J P Hubbard. NestedMICA as an ab initio protein motif discovery tool. *BMC bioinformatics*, 9(1):1–12, 2008.
- Viacheslav V Dolgikh. Activities of enzymes of carbohydrate and energy metabolism of the intracellular stages of the microsporidian, *Nosema grylli*. Technical Report 3, 2000.
- Viacheslav V. Dolgikh, Igor V. Senderskiy, Olga A. Pavlova, Anton M. Naumov, and Galina V. Beznoussenko. Immunolocalization of an alternative respiratory chain in antonospora (*Paranosema*) locustae spores: Mitosomes retain their role in microsporidial energy metabolism. *Eukaryotic Cell*, 10(4), 2011. ISSN 15359778. doi: 10.1128/EC.00283-10.

- Yuemei Dong, Fabio Manfredini, and George Dimopoulos. Implication of the mosquito midgut microbiota in the defense against malaria parasites. *PLoS Pathogens*, 5(5), 2009. ISSN 15537366. doi: 10.1371/journal.ppat.1000423.
- Sean R Eddy. HMMER User ' s Guide: Biological sequence analysis using profile hidden Markov models. 2020.
- Robert C. Edgar, Roque Moraes Drive, and Mill Valley. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004. ISSN 03051048. doi: 10.1093/nar/gkh340.
- Edward Eisenstein, Gary L. Gilliland, Osnat Herzberg, John Moulton, John Orban, Roberto J. Poljak, Linda Banerjee, Delwood Richardson, and Andrew J. Howard. Biological function made crystal clear - Annotation of hypothetical proteins via structural genomics. *Current Opinion in Biotechnology*, 11(1):25–30, 2000. ISSN 09581669. doi: 10.1016/S0958-1669(99)00063-4.
- E. Elayaraja, K. Thangavel, B. Ramya, and M. Chitraregha. Extraction of motif patterns from protein sequence using Rough-K-Means algorithm. In *Procedia Engineering*, volume 30, pages 814–820, 2012. doi: 10.1016/j.proeng.2012.01.932.
- David M. Emms and Steven Kelly. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1):1–14, 2015. ISSN 1474760X. doi: 10.1186/s13059-015-0721-2. URL <http://dx.doi.org/10.1186/s13059-015-0721-2>.
- David M. Emms and Steven Kelly. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1):1–14, 2019. ISSN 1474760X. doi: 10.1186/s13059-019-1832-y.
- Philipp Engel and Nancy A. Moran. The gut microbiota of insects - diversity in structure and function. *FEMS Microbiology Reviews*, 37(5):699–735, 2013. ISSN 01686445. doi: 10.1111/1574-6976.12025.
- Who-coordinated Multi-country Evaluation. WHO - Implications of insecticide resistance for malaria vector control. (November), 2016.

- Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016. ISSN 14602059. doi: 10.1093/bioinformatics/btw354.
- Naomi M Fast and Patrick J Keeling. The Fungal Roots of Microsporidian Parasites. In *Insect-Fungal Associations Ecology and Evolution*. Oxford University Press, 2 2005. ISBN 9780195166521. doi: 10.1093/oso/9780195166521.003.0005. URL <https://doi.org/10.1093/oso/9780195166521.003.0005>.
- Ngonidzashe Faya, David L Penkler, and Özlem Tastan Bishop. Human , vector and parasite Hsp90 proteins : A comparative bioinformatics analysis. *FEBS OPEN BIO*, 5:916–927, 2015. ISSN 2211-5463. doi: 10.1016/j.fob.2015.11.003. URL <http://dx.doi.org/10.1016/j.fob.2015.11.003>.
- Robert D Finn, Jody Clements, and Sean R Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, 39(suppl_2): W29–W37, 2011.
- Jullien M. Flynn, Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, 117(17):9451–9457, 4 2020. ISSN 1091-6490 (Electronic). doi: 10.1073/pnas.1921046117.
- National Center for Biotechnology Information (US) and Christiam Camacho. *BLAST (r) command line applications user manual*. National Center for Biotechnology Information (US) Bethesda, MD, 2008.
- Caspar Franzen and Bernd Salzberger. Analysis of the β -tubulin gene from *Vittaforma corneae* suggests benzimidazole resistance. *Antimicrobial Agents and Chemotherapy*, 52(2):790–793, 2 2008. ISSN 00664804. doi: 10.1128/AAC.00928-07.
- Frances V. Fuller-Pace. DExD/H box RNA helicases: Multifunctional proteins with important roles in transcriptional regulation. *Nucleic Acids Research*, 34(15):4206–4215, 2006. ISSN 03051048. doi: 10.1093/nar/gkl460.

- M. Y. Galperin. Conserved 'hypothetical' proteins: New hints and new puzzles. *Comparative and Functional Genomics*, 2(1):14–18, 2001. ISSN 15316912. doi: 10.1002/cfg.66.
- Han Gao, Chunlai Cui, Lili Wang, Marcelo Jacobs-Lorena, and Sibao Wang. Mosquito Microbiota and Implications for Disease Control. *Trends in Parasitology*, 36(2):98–111, 2020. ISSN 14715007. doi: 10.1016/j.pt.2019.12.001. URL <https://doi.org/10.1016/j.pt.2019.12.001>.
- Juan Antonio Raygoza Garay, Grant L. Hughes, Vikas Koundal, Jason L. Rasgon, and Michael M. Mwangi. Genome sequence of *Elizabethkingia anophelis* strain EaAs1, isolated from the Asian malaria mosquito *Anopheles stephensi*. *Genome Announcements*, 4(2):15–16, 2016. ISSN 21698287. doi: 10.1128/genomeA.00084-16.
- Arkadiy I. Garber, Catherine R. Armbruster, Stella E. Lee, Vaughn S. Cooper, Jennifer M. Bomberger, and Sean M. McAllister. SprayNPray: user-friendly taxonomic profiling of genome and metagenome contigs. *BMC Genomics*, 23(1):1–8, 2022. ISSN 14712164. doi: 10.1186/s12864-022-08382-2. URL <https://doi.org/10.1186/s12864-022-08382-2>.
- Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, S'everine Duvaud, Marc R Wilkins, Ron D Appel, and Amos Bairoch. *Protein identification and analysis tools on the ExPASy server*. Springer, 2005a.
- Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, Séverine Duvaud, Marc R Wilkins, Ron D Appel, and Amos Bairoch. The Proteomics Protocols Handbook. *The Proteomics Protocols Handbook*, pages 571–608, 2005b. doi: 10.1385/1592598900.
- Mathilde Gendrin, Faye H. Rodgers, Rakiswendé S. Yerbanga, Jean Bosco Ouédraogo, María-Gloria Basáñez, Anna Cohuet, and George K. Christophides. Antibiotics in ingested human blood affect the mosquito microbiota and capacity to transmit malaria. *Nature Communications*, 6, 2015. ISSN 2041-1723. doi: 10.1038/ncomms6921.

- Agnès Germot, Hervé Philippe, and Hervé Le Guyader. Evidence for loss of mitochondria in microsporidia from a mitochondrial-type HSP70 in *Nosema locustae*. *Molecular and Biochemical Parasitology*, 87(2):159–168, 1997. ISSN 01666851. doi: 10.1016/S0166-6851(97)00064-9.
- Erin E. Gill, James J. Becnel, and Naomi M. Fast. ESTs from the microsporidian *Edhazardia aedis*. *BMC Genomics*, 9, 6 2008. ISSN 14712164. doi: 10.1186/1471-2164-9-296.
- Janice I Glasgow, Evan W Steeg, and Suzanne Fortier. Motif Discovery in Protein Structure Databases., 1999.
- Dörte Goertz and Gernot Hoch. Vertical transmission and overwintering of microsporidia in the gypsy moth, *Lymantria dispar*. *Journal of Invertebrate Pathology*, 99(1):43–48, 2008. ISSN 00222011. doi: 10.1016/j.jip.2008.03.008.
- Alina V. Goldberg, Sabine Molik, Anastasios D. Tsaousis, Karina Neumann, Grit Kuhnke, Frederic Delbac, Christian P. Vivares, Robert P. Hirt, Roland Lill, and T. Martin Embley. Localization and functionality of microsporidian iron-sulphur cluster assembly proteins. *Nature*, 452(7187):624–628, 4 2008. ISSN 14764687. doi: 10.1038/nature06606.
- Michael Groll, Monica Bajorek, Alwin Köhler, Luis Moroder, David M Rubin, Robert Huber, Michael H Glickman, and Daniel Finley. A gated channel into the proteasome core particle. 7(11):1062–1067, 2000.
- Rui Guo, Guangli Cao, Yahong Lu, Renyu Xue, Dhiraj Kumar, Xiaolong Hu, and Chengliang Gong. Exogenous gene can be integrated into *Nosema bombycis* genome by mediating with a non-transposon vector. *Parasitology Research*, 115(8):3093–3098, 2016. ISSN 14321955. doi: 10.1007/s00436-016-5064-8. URL <http://dx.doi.org/10.1007/s00436-016-5064-8>.
- Akanksha Gupta, Pallavi Mohanty, and Sonika Bhatnagar. Protein Structure Prediction using Homology Modeling: Methods and Tools. In *Methods and Algorithms for Molecular Docking-Based Drug Design and Discovery*, pages 339–359. IGI Global, 2016.

Ankit Gupta, Rasna Gupta, and Ram Lakhan Singh. Microbes and environment. *Principles and applications of environmental biotechnology for a sustainable future*, pages 43–84, 2017.

Saurabh Gupta, Yashbir Singh, Himansu Kumar, Utkarsh Raj, A. R. Rao, and Pritish Kumar Varadwaj. Identification of Novel Abiotic Stress Proteins in *Triticum aestivum* Through Functional Annotation of Hypothetical Proteins. *Interdisciplinary Sciences: Computational Life Sciences*, 10(1):205–220, 2018. ISSN 18671462. doi: 10.1007/s12539-016-0178-3.

Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUASt : quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013. doi: 10.1093/bioinformatics/btt086.

Karen L Haag, Timothy Y James, Jean-françois Pombert, Ronny Larsson, and Tobias M M Schaer. Evolution of a morphological novelty occurred before genome compaction in a lineage of extreme parasites. 111(43), 2014. doi: 10.1073/pnas.1410442111.

Karen L. Haag, Jean François Pombert, Yukun Sun, Nathalia Rammé M. De Albuquerque, Brendan Batliner, Peter Fields, Tiago Falcon Lopes, and Dieter Ebert. Microsporidia with Vertical Transmission Were Likely Shaped by Non-adaptive Processes. *Genome Biology and Evolution*, 12(1):3599–3614, 2019. ISSN 17596653. doi: 10.1093/gbe/evz270.

Karen Luisa Haag and Nathalia Rammé Medeiros De Albuquerque. Using average nucleotide identity (ANI) to evaluate microsporidia species boundaries based on their genetic relatedness. *The Journal of eukaryotic microbiology*, 70, 9 2022. doi: 10.1111/jeu.12944. URL <https://discovery.researcher.life/article/using-average-nucleotide-identity-ani-to-evaluate-microsporidia-species-boundaries-based-on-bffe7030c2233d08a9b2bfc03418c5e0>.

Jeppe Hallgren, Konstantinos D Tsirigos, Mads Damgaard Pedersen, José Juan, Almagro Armenteros, Paolo Marcatili, Henrik Nielsen, Anders Krogh, and Ole Winther. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv*, page 2022.04.08.487609,

2022. URL <https://www.biorxiv.org/content/10.1101/2022.04.08.487609v1%0Ahttps://www.biorxiv.org/content/10.1101/2022.04.08.487609v1.abstract>.
- L. Y. Han, C. Z. Cai, Z. L. Ji, Z. W. Cao, J. Cui, and Y. Z. Chen. Predicting functional family of novel enzymes irrespective of sequence similarity: A statistical learning approach. *Nucleic Acids Research*, 32(21):6437–6444, 2004. ISSN 03051048. doi: 10.1093/nar/gkh984.
- Rowan Hatherley, David K. Brown, Michael Glenister, and Özlem Tastan Bishop. PRIMO: An interactive homology modeling pipeline. *PLoS ONE*, 11(11):1–20, 2016. ISSN 19326203. doi: 10.1371/journal.pone.0166698.
- Qiang He, Charles R. Vossbrinck, Qiong Yang, Xian Zhi Meng, Jian Luo, Guo Qing Pan, Ze Yang Zhou, and Tian Li. Evolutionary and functional studies on microsporidian ATP-binding cassettes: Insights into the adaptation of microsporidia to obligated intracellular parasitism. *Infection, Genetics and Evolution*, 68(December 2018):136–144, 2019. ISSN 15677257. doi: 10.1016/j.meegid.2018.12.022. URL <https://doi.org/10.1016/j.meegid.2018.12.022>.
- Shivanand Hegde, Jason L. Rasgon, and Grant L. Hughes. The microbiome modulates arbovirus transmission in mosquitoes. *Current Opinion in Virology*, 15, 2015. ISSN 18796265. doi: 10.1016/j.coviro.2015.08.011.
- Eva Heinz, Tom A. Williams, Sirintra Nakjang, Christophe J. Noël, Daniel C. Swan, Alina V. Goldberg, Simon R. Harris, Thomas Weinmaier, Stephanie Markert, Dörte Becher, Jörg Bernhardt, Tal Dagan, Christian Hacker, John M. Lucocq, Thomas Schweder, Thomas Rattei, Neil Hall, Robert P. Hirt, and T. Martin Embley. The Genome of the Obligate Intracellular Parasite *Trachipleistophora hominis*: New Insights into Microsporidian Genome Dynamics and Reductive Evolution. *PLoS Pathogens*, 8(10), 2012. ISSN 15537366. doi: 10.1371/journal.ppat.1002979.
- Eva Heinz, Christian Hacker, Paul Dean, John Mifsud, Alina V. Goldberg, Tom A. Williams, Sirintra Nakjang, Alison Gregory, Robert P. Hirt, John M. Lucocq, Edmund R.S. S Kunji, and T. Martin Embley. Plasma Membrane-Located

- Purine Nucleotide Transport Proteins Are Key Components for Host Exploitation by Microsporidian Intracellular Parasites. *PLoS Pathogens*, 10(12):1–12, 2014. ISSN 15537374. doi: 10.1371/journal.ppat.1004547.
- Janet Hemingway. The role of vector control in stopping the transmission of malaria: threats and opportunities. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369(1645):20130431, 2014. ISSN 1471-2970. doi: 10.1098/rstb.2013.0431. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4024224&tool=pmcentrez&rendertype=abstract>.
- Jeremy K. Herren and Bruno Lemaitre. Spiroplasma and host immunity: Activation of humoral immune responses increases endosymbiont load and susceptibility to certain Gram-negative bacterial pathogens in *Drosophila melanogaster*. *Cellular Microbiology*, 13(9):1385–1396, 2011. ISSN 14625814. doi: 10.1111/j.1462-5822.2011.01627.x.
- J.K. Jeremy K. Herren, Lilian Mbaisi, Enock Mararo, E.E. Edward E. Makhulu, V.A. Victor A. Mobegi, Hellen Butungi, M.V. Mancini, J.W. Joseph W. Oundo, Evan E.T. Teal, Silvain Pinaud, Mara K.N. M.K.N. Lawniczak, Jordan Jabara, Godfrey Nattoh, and S.P. Steven P. Sinkins. A microsporidian impairs *Plasmodium falciparum* transmission in *Anopheles arabiensis* mosquitoes. *Nature Communications*, 11(1), 2020. ISSN 20411723. doi: 10.1038/s41467-020-16121-y. URL <http://dx.doi.org/10.1038/s41467-020-16121-y>.
- Mariano Higes, Raquel Martín-Hernández, Cristina Botías, Encarna Garrido Bailón, Amelia V. González-Porto, Laura Barrios, M. Jesús Del Nozal, José L. Bernal, Juan J. Jiménez, Pilar García Palencia, and Aránzazu Meana. How natural infection by *Nosema ceranae* causes honeybee colony collapse. *Environmental Microbiology*, 10(10):2659–2669, 2008. ISSN 14622912. doi: 10.1111/j.1462-2920.2008.01687.x.
- Christopher F Higgins. ScienceDirect.com - Research in Microbiology - ABC transporters: physiology, structure and mechanism – an overview. *Research in Microbiology*, 152(3-4):205–210, 2001. ISSN

09232508. URL [http://ntserver1.wsulibs.wsu.edu:2091/science/article/pii/S0923250801011937%5Cnhttp://dx.doi.org/10.1016/S0923-2508\(01\)01193-7](http://ntserver1.wsulibs.wsu.edu:2091/science/article/pii/S0923250801011937%5Cnhttp://dx.doi.org/10.1016/S0923-2508(01)01193-7).
- Qiang Huang. Evolution of Dicer and Argonaute orthologs in microsporidian parasites. *Infection, Genetics and Evolution*, 65(August):329–332, 2018. ISSN 15677257. doi: 10.1016/j.meegid.2018.08.011. URL <https://doi.org/10.1016/j.meegid.2018.08.011>.
- Qiang Huang, Zhi Hao Wu, Wen Feng Li, Rui Guo, Jin Shan Xu, Xiao Qun Dang, Zheng Gang Ma, Yan Ping Chen, and Jay D. Evans. Genome and Evolutionary Analysis of *Nosema ceranae*: A Microsporidian Parasite of Honey Bees. *Frontiers in Microbiology*, 12(June):1–17, 2021. ISSN 1664302X. doi: 10.3389/fmicb.2021.645353.
- Qiang Huang, Hamed Mirjalali, and Zeyang Zhou. Editorial: Genomic and evolutionary analysis of microsporidian parasites. *Frontiers in Microbiology*, 14(1):1–2, 2023. ISSN 1664-302X. doi: 10.3389/fmicb.2023.1248661. URL <https://doi.org/10.3389/fmicb.2023.1248661>.
- Daniel H. Huson, Daniel C. Richter, Christian Rausch, Tobias Dezulian, Markus Franz, and Regula Rupp. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, 8:1–6, 2007. ISSN 14712105. doi: 10.1186/1471-2105-8-460.
- Md. Saiful Islam, Shah Md. Shahik, Md. Sohel, Noman I. A. Patwary, and Md. Anayet Hasan. In Silico Structural and Functional Annotation of Hypothetical Proteins of *Vibrio cholerae* O139. *Genomics & Informatics*, 13(2):53, 10 2015. ISSN 2234-0742. doi: 10.5808/GI.2015.13.2.53. URL <http://genominfo.org/journal/view.php?doi=10.5808/GI.2015.13.2.53>.
- Mohamed Issa and Aboul Ella Hassanien. Multiple sequence alignment optimization using meta-heuristic techniques. In *Data Analytics in Medicine: Concepts, Methodologies, Tools, and Applications*, pages 565–579. IGI Global, 2020.
- S. Iuchi. Three classes of C2H2 zinc finger proteins. *Cellular and Molecular Life Sciences*, 58(4):625–635, 2001. ISSN 1420682X. doi: 10.1007/PL00000885.

- Péter Jacsó. Google scholar: The pros and the cons. *Online Information Review*, 29(2):208–214, 2005. ISSN 14684527. doi: 10.1108/14684520510598066.
- Buddhi Prakash Jain and Shweta Pandey. WD40 Repeat Proteins: Signalling Scaffold with Diverse Functions. *Protein Journal*, 37(5):391–406, 2018. ISSN 15734943. doi: 10.1007/s10930-018-9785-7. URL <http://dx.doi.org/10.1007/s10930-018-9785-7>.
- Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. *Genome biology*, 17:1–11, 2016.
- Timothy Y. James, Adrian Pelin, Linda Bonen, Steven Ahrendt, Divya Sain, Nicolas Corradi, and Jason E. Stajich. Shared signatures of parasitism and phylogenomics unite cryptomycota and microsporidia. *Current Biology*, 23(16):1548–1553, 2013. ISSN 09609822. doi: 10.1016/j.cub.2013.06.057. URL <http://dx.doi.org/10.1016/j.cub.2013.06.057>.
- Rolf Jessberger. The many functions of SMC proteins in chromosome dynamics. *Nature Reviews Molecular Cell Biology*, 3(10):767–778, 2002.
- Mark Johnson, Irena Zaretskaya, Yan Raytselis, Yuri Merezuk, Scott McGinnis, and Thomas L. Madden. NCBI BLAST: a better web interface. *Nucleic acids research*, 36(Web Server issue):5–9, 2008. ISSN 13624962. doi: 10.1093/nar/gkn201.
- Natapong Jupatanakul, Jutharat Pengon, Shiela Marie Gines Selisana, Waeowalee Choksawangkar, Nongluck Jaito, Atiporn Saeung, Ratchanu Bunyong, Nava-porn Posayapisit, Khongkhwan Thammattinna, Nuttiya Kalpongkul, Kittipat Aupalee, Trairak Pisitkun, and Sumalee Kamchonwongpaisan. *Serratia marcescens* secretes proteases and chitinases with larvicidal activity against *Anopheles dirus*. *Acta Tropica*, 212(August):105686, 2020. ISSN 18736254. doi: 10.1016/j.actatropica.2020.105686. URL <https://doi.org/10.1016/j.actatropica.2020.105686>.

- Wolfgang Kabsch and Joël Vandekerckhove. Structure and function of actin. *Annual Review of Biophysics and Biomolecular Structure*, 21(103):49–76, 1992. ISSN 10568700. doi: 10.1146/annurev.bb.21.060192.000405.
- Lukas Käll, Anders Krogh, and Erik L.L. Sonnhammer. A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, 338(5):1027–1036, 2004. ISSN 00222836. doi: 10.1016/j.jmb.2004.03.016.
- Subha Kalyaanamoorthy, Bui Quang Minh, Thomas K.F. Wong, Arndt Von Haeseler, and Lars S. Jermiin. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589, 2017. ISSN 15487105. doi: 10.1038/nmeth.4285.
- Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, 2016. ISSN 13624962. doi: 10.1093/nar/gkv1070.
- Eli Levy Karin, Milot Mirdita, and Johannes Söding. MetaEuk-sensitive, high-throughput gene discovery and annotation for large-scale eukaryotic metagenomics Unicellular eukaryotes are diverse but under studied MetaEuk discovers protein-coding genes in eukaryotic metagenomic data Benchmark Input 1. Putative protein fragment translation. Technical report.
- Michaël D. Katinka, Simone Duprat, Emmanuel Cornillott, Guy Méténler, Fabienne Thomarat, Gérard Prensier, Valérie Barbe, Eric Peyretailade, Philippe Brottier, Patrick Wincker, Frédéric Delbac, Hicham El Alaoul, Pierre Peyret, William Saurin, Manolo Gouy, and Jean Weissenbach. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature*, 414(6862):450–453, 2001. ISSN 00280836. doi: 10.1038/35106579.
- Matthew Kearse, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, Alex Cooper, Sidney Markowitz, Chris Duran, Tobias Thierer, Bruce Ashton, Peter Meintjes, and Alexei Drummond. Geneious Basic: An integrated and extendable desktop software platform for the

- organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts199.
- Patrick J. Keeling and Nicolas Corradi. Balancing loss of function with shrinking genomes in the microsporidia. *Virulence*, 2(1):67–70, 2011. ISSN 21505608. doi: 10.4161/viru.2.1.14606.
- Patrick J. Keeling and Naomi M. Fast. Microsporidia: Biology and evolution of highly reduced intracellular parasites. *Annual Review of Microbiology*, 56: 93–116, 2002. ISSN 00664227. doi: 10.1146/annurev.micro.56.012302.160854.
- Patrick J. Keeling and Jeffrey D. Palmer. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618, 2008. ISSN 14710056. doi: 10.1038/nrg2386.
- Patrick J. Keeling and Claudio H. Slamovits. Simplicity and complexity of microsporidian genomes, 12 2004. ISSN 15359778.
- Patrick J. Keeling, Naomi M. Fast, Joyce S. Law, Bryony A.P. Williams, and Claudio H. Slamovits. Comparative genomics of microsporidia. *Folia Parasitologica*, 52(1-2):8–14, 2005. ISSN 00155683. doi: 10.14411/fp.2005.002.
- Patrick J. Keeling, Nicolas Corradi, Hilary G. Morrison, Karen L. Haag, Dieter Ebert, Louis M. Weiss, Donna E. Akiyoshi, and Saul Tzipori. The reduced genome of the parasitic microsporidian *Enterocytozoon bieneusi* lacks genes for core carbon metabolism. *Genome Biology and Evolution*, 2(1):304–309, 2010. ISSN 17596653. doi: 10.1093/gbe/evq022.
- Ozlem Keskin, Nurcan Tuncbag, and Attila Gursoy. Predicting protein–protein interactions from the molecular to the proteome level. *Chemical reviews*, 116(8):4884–4909, 2016.
- Manfred Koegl, Thorsten Hoppe, Stephan Schlenker, Helle D. Ulrich, Thomas U. Mayer, and Stefan Jentsch. A novel ubiquitination factor, E4, is involved in multiubiquitin chain assembly. *Cell*, 96(5):635–644, 1999. ISSN 00928674. doi: 10.1016/S0092-8674(00)80574-7.

- J C Koella and P Agnew. Blood-feeding success of the mosquito *Aedes aegypti* depends on the transmission route of its parasite *Edhazardia aedis*, 1997. ISSN 0030-1299.
- Jurgen Kopp and Torsten Schwede. Automated protein structure homology modeling: a progress report. *Pharmacogenomics*, 5(4):405–416, 2004.
- Paula Korkuc and Dirk Walther. Physicochemical characteristics of structurally determined metabolite-protein and drug-protein binding events with respect to binding specificity. *Frontiers in Molecular Biosciences*, 2(SEP):1–20, 2015. ISSN 2296889X. doi: 10.3389/fmolb.2015.00051.
- Evgenia V. Kriventseva, Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A. Simão, and Evgeny M. Zdobnov. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, 47(D1):D807–D811, 2019. ISSN 13624962. doi: 10.1093/nar/gky1053.
- Kundan Kumar, Amresh Prakash, Munazzah Tasleem, Asimul Islam, Faizan Ahmad, and Md Imtaiyaz Hassan. Functional annotation of putative hypothetical proteins from *Candida dubliniensis*. *Gene*, 543(1):93–100, 2014. ISSN 18790038. doi: 10.1016/j.gene.2014.03.060. URL <http://dx.doi.org/10.1016/j.gene.2014.03.060>.
- Sudhir Kumar, Glen Stecher, Daniel Peterson, and Koichiro Tamura. MEGA-CC: Computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics*, 28(20):2685–2686, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts507.
- Sudhir Kumar, Glen Stecher, Michael Li, Christina Knyaz, and Koichiro Tamura. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, 35(6):1547–1549, 2018. ISSN 15371719. doi: 10.1093/molbev/msy096.
- A. H.Y. Kwan, R. D. Winefield, M. Sunde, J. M. Matthews, R. G. Haverkamp, M. D. Templeton, and J. P. Mackay. Structural basis for rodlet assembly in

- fungus hydrophobins. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3621–3626, 2006. ISSN 00278424. doi: 10.1073/pnas.0505704103.
- Dominik R Laetsch, Mark L Blaxter, and Richard M Leggett. BlobTools : Interrogation of genome assemblies [version 1 ; peer review : 2 approved with reservations]. 1287:1–16, 2020.
- Chien Yueh Lee, Yu Chiao Chiu, Liang Bo Wang, Yu Lun Kuo, Eric Y. Chuang, Liang Chuan Lai, and Mong Hsun Tsai. Common applications of next-generation sequencing technologies in genomic research, 2 2013. ISSN 22196803.
- Soo Chan Lee, Joseph Heitman, and Joseph E. Ironside. Sex and the Microsporidia. *Microsporidia: Pathogens of Opportunity: First Edition*, (1983): 231–243, 2014. doi: 10.1002/9781118395264.ch8.
- Jonathan G. Lees, David Lee, Romain A. Studer, Natalie L. Dawson, Ian Sillitoe, Sayoni Das, Corin Yeats, Benoit H. Dessailly, Robert Rentzsch, and Christine A. Orengo. Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis. *Nucleic Acids Research*, 42(D1):240–245, 2014. ISSN 03051048. doi: 10.1093/nar/gkt1205.
- Jordan Leroy, Marjorie Cornu, Anne-Sophie Deleplancque, Aldert Bart, Severine Loridant, Emilie Frealle, Emmanuel Dutoit, Olivier Gaillot, Tom Van Gool, François Puisieux, et al. Case report: ocular microsporidiosis: case in a patient returning from india and review of the literature. *The American Journal of Tropical Medicine and Hygiene*, 99(1):90, 2018.
- Ivica Letunic, Supriya Khedkar, and Peer Bork. SMART: Recent updates, new developments and status in 2020. *Nucleic Acids Research*, 49(D1):D458–D460, 2021. ISSN 13624962. doi: 10.1093/nar/gkaa937.
- Eli Levy Karin, Milot Mirdita, and Johannes Söding. MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome*, 8(1), 4 2020. ISSN 20492618. doi: 10.1186/s40168-020-00808-x.

- Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp324.
- Zhi Li, Youjin Hao, Linling Wang, Heng Xiang, and Zeyang Zhou. Genome-wide identification and comprehensive analyses of the kinomes in four pathogenic microsporidia species. *PLoS ONE*, 9(12):1–27, 2014. ISSN 19326203. doi: 10.1371/journal.pone.0115890.
- Roland Lill, Rafal Dutkiewicz, Hans-Peter Elsässer, Anja Hausmann, Daili J A Netz, Antonio J Pierik, Oliver Stehling, Eugen Urzica, and Ulrich Mühlenhoff. Mechanisms of iron-sulfur protein maturation in mitochondria, cytosol and nucleus of eukaryotes. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1763(7):652–667, 2006.
- Weiqiang Lin, Mian Wang, Hui Jin, and Hong-Guo Yu. Cohesin plays a dual role in gene regulation and sister-chromatid cohesion during meiosis in *Saccharomyces cerevisiae*. *Genetics*, 187(4):1041–1051, 2011.
- Jiaquan Liu, Jong-Bong Lee, and Richard Fishel. Stochastic processes and component plasticity governing DNA mismatch repair. *Journal of molecular biology*, 430(22):4456–4468, 2018.
- Wei Liu, Ziyi Wang, Ronghui You, Chenghan Xie, Hong Wei, Yi Xiong, Jianyi Yang, and Shanfeng Zhu. Protein language model powers accurate and fast sequence search for remote homology. *bioRxiv*, pages 2023–2024, 2023.
- C Loubès. Meiosis in Microsporidia: effects on biological cycles. *The Journal of Protozoology*, 26(2):200–208, 1979.
- Jennifer Lu and Steven L. Salzberg. Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. *Microbiome*, 8(1):1–11, 2020. ISSN 20492618. doi: 10.1186/s40168-020-00900-2.
- P. Lukasik, H. Guo, M. Van Asch, J. Ferrari, and H. C J Godfray. Protection against a fungal pathogen conferred by the aphid facultative endosymbionts *Rickettsia* and *Spiroplasma* is expressed in multiple host genotypes and

- species and is not influenced by co-infection with another symbiont. *Journal of Evolutionary Biology*, 26(12):2654–2661, 2013. ISSN 1010061X. doi: 10.1111/jeb.12260.
- Hongmei Luo and Shilin Chen. Progress on the Transcriptome Analysis of Medicinal Plants with Next-Generation Sequencing Technologies. In *Encyclopedia of Analytical Chemistry*, pages 1–21. Wiley, 3 2019. doi: 10.1002/9780470027318.a9936.pub2.
- Jian Luo, Qiang He, Jin Zhi Xu, Chen Xu, Yin Ze Han, Hai Long Gao, Xian Zhi Meng, Guo Qing Pan, Tian Li, and Ze Yang Zhou. Microsporidia infection upregulates host energy metabolism but maintains ATP homeostasis. *Journal of Invertebrate Pathology*, 186(April):107596, 2021. ISSN 10960805. doi: 10.1016/j.jip.2021.107596. URL <https://doi.org/10.1016/j.jip.2021.107596>.
- Roland Luthy, James Bowei, and David Einsenberg. Verify3D: Assessment of protein models with three-dimensional profiles. *Methods in enzymology*, 277: 396–404, 1997.
- L. Mahadevan, C. S. Riera, and Jennifer H. Shin. Structural dynamics of an actin spring. *Biophysical Journal*, 100(4):839–844, 2011. ISSN 15420086. doi: 10.1016/j.bpj.2010.12.3743. URL <http://dx.doi.org/10.1016/j.bpj.2010.12.3743>.
- Edward E. Makhulu, Thomas O. Onchuru, Joseph Gichuhi, Fidel G. Otieno, Anne W. Wairimu, Joseph N. Muthoni, Lizette Koekemoer, and Jeremy K. Herren. Localization and tissue tropism of the symbiont Microsporidia MB in the germ line and somatic tissues of *Anopheles arabiensis*. *mBio*, 1 2023. ISSN 21507511. doi: 10.1128/mbio.02192-23.
- Shehre-Banoo Malik, Arthur W Pightling, Lauren M Stefaniak, Andrew M Schurko, and John M Logsdon Jr. An expanded inventory of conserved meiotic genes provides evidence for sex in *Trichomonas vaginalis*. *PloS one*, 3(8):e2879, 2008.
- Mosè Manni, Matthew R. Berkeley, Mathieu Seppey, Felipe A. Simão, and Evgeny M. Zdobnov. BUSCO Update: Novel and Streamlined Workflows along

- with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, 38(10): 4647–4654, 2021. ISSN 15371719. doi: 10.1093/molbev/msab199.
- Aron Marchler-Bauer, Yu Bo, Lianyi Han, Jane He, Christopher J. Lanczycki, Shennan Lu, Farideh Chitsaz, Myra K. Derbyshire, Renata C. Geer, Noreen R. Gonzales, Marc Gwadz, David I. Hurwitz, Fu Lu, Gabriele H. Marchler, James S. Song, Narmada Thanki, Zhouxi Wang, Roxanne A. Yamashita, Dachuan Zhang, Chanjuan Zheng, Lewis Y. Geer, and Stephen H. Bryant. CD-D/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, 45(D1):D200–D203, 2017. ISSN 13624962. doi: 10.1093/nar/gkw1129.
- Mindaugas Margelevičius and Česlovas Venclovas. Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. *BMC bioinformatics*, 11(1):1–14, 2010.
- Marc A Marti-Renom, Bozidar Yerkovich, and Andrej Sali. Comparative protein structure prediction. *Current Protocols in Protein Science*, 28(1):2–9, 2002.
- Marc A Marti-Renom, M S Madhusudhan, Narayanan Eswar, Ursula Pieper, Minyi Shen, Andrej Sali, Andras Fiser, Nebojsa Mirkovic, Bino John, and Ashley Stuart. Modeling protein structure from its sequence. *Current Protocols in Bioinformatics*, 3(1):1–5, 2003.
- Anne Caroline Mascarenhas dos Santos, Alexander Thomas Julian, and Jean-François Pombert. The rad9–rad1–hus1 dna repair clamp is found in microsporidia. *Genome Biology and Evolution*, 14(4):evac053, 2022.
- Gaston K. Mazandu and Nicola J. Mulder. Function prediction and analysis of mycobacterium tuberculosis hypothetical proteins. *International Journal of Molecular Sciences*, 13(6):7283–7302, 2012. ISSN 14220067. doi: 10.3390/ijms13067283.
- Elizabeth a McGraw and Scott L O’Neill. Beyond insecticides: new thinking on an ancient problem. *Nature reviews. Microbiology*, 11(3):181–93, 2013. ISSN 1740-

1534. doi: 10.1038/nrmicro2968. URL <http://www.ncbi.nlm.nih.gov/pubmed/23411863>(Accessed24/04/2016).

Muhammad Aamer Mehmood, Ujala Sehar, and Niaz Ahmad. Use of bioinformatics tools in different spheres of life sciences. *Journal of Data Mining in Genomics & Proteomics*, 5(2):1, 2014.

Sergey V. Melnikov, Kasidet Manakongtreecheep, Keith D. Rivera, Arthur Makarenko, Darryl J. Pappin, and Dieter Söll. Muller’s Ratchet and ribosome degeneration in the obligate intracellular parasites Microsporidia. *International Journal of Molecular Sciences*, 19(12):1–12, 2018. ISSN 14220067. doi: 10.3390/ijms19124125.

Vesna Memišević, Memi Memišević, Tijana Milenković, Nataša Nataš, and Nataša Pržulj. Complementarity of network and sequence information in homologous proteins. 2010. doi: 10.2390/biecoll-jib-2010-135. URL <http://journal.imbio.de>.

Xian Zhi Meng, Bo Luo, Xiang You Tang, Qiang He, Ting Rong Xiong, Zhuo Ya Fang, Guoqing Pan, Tian Li, and Ze Yang Zhou. Pathological analysis of silkworm infected by two microsporidia *Nosema bombycis* CQ1 and *Vairimorpha necatrix* BM. *Journal of Invertebrate Pathology*, 153(December 2017): 75–84, 2018. ISSN 10960805. doi: 10.1016/j.jip.2017.12.005. URL <https://doi.org/10.1016/j.jip.2017.12.005>.

Charlène N.T. Mfangnia, Henri E.Z. Tonnang, Berge Tsanou, and Jeremy Herren. Mathematical modelling of the interactive dynamics of wild and Microsporidia MB-infected mosquitoes. *Mathematical Biosciences and Engineering*, 20(8): 15167–15200, 2023. ISSN 15510018. doi: 10.3934/mbe.2023679.

Huaiyu Mi, Dustin Ebert, Anushya Muruganujan, Caitlin Mills, Laurent Philippe Albou, Tremayne Mushayamaha, and Paul D. Thomas. PANTHER version 16: A revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research*, 49(D1):D394–D403, 2021. ISSN 13624962. doi: 10.1093/nar/gkaa1106.

Guillaume Minard, Patrick Mavingui, and Claire Valiente Moro. Diversity and

- function of bacterial microbiota in the mosquito holobiont. *Parasites & Vectors*, 6, 2013. ISSN 1756-3305. doi: 10.1186/1756-3305-6-146.
- Ryuhei Minei, Ryo Hoshina, and Atsushi Ogura. De novo assembly of middle-sized genome using MinION and Illumina sequencers 06 Biological Sciences 0604 Genetics. *BMC Genomics*, 19(1):1–12, 2018. ISSN 14712164. doi: 10.1186/s12864-018-5067-1.
- Dengming Ming, Min Han, and Xiongbo An. Sequence-based prediction of function site and protein-ligand interaction by a functionally annotated domain profile database. *arXiv preprint arXiv:1701.08086*, 2017.
- Diego Miranda-Saavedra, Michael J.R. Stark, Jeremy C. Packer, Christian P. Vives, Christian Doerig, and Geoffrey J. Barton. The complement of protein kinases of the microsporidium *Encephalitozoon cuniculi* in relation to those of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. *BMC Genomics*, 8: 1–21, 2007. ISSN 14712164. doi: 10.1186/1471-2164-8-309.
- Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L.L. Sonnhammer, Silvio C.E. Tosatto, Lisanna Paladin, Shriya Raj, Lorna J. Richardson, Robert D. Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 2021. ISSN 13624962. doi: 10.1093/nar/gkaa913.
- Derek Mittleider, Linda C Green, Victoria H Mann, Scott F Michael, Elizabeth S Didier, and Paul J Brindley. Sequence survey of the genome of the opportunistic microsporidian pathogen, *Vittaforma corneae*. *Journal of Eukaryotic Microbiology*, 49(5):393–401, 2002.
- Steffen Möller, Michael D.R. Croning, and Rolf Apweiler. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, 17(7):646–653, 2001. ISSN 13674803. doi: 10.1093/bioinformatics/17.7.646.
- Catherine Mooney, Norman Davey, Alberto J M Martin, Ian Walsh, Denis C Shields, and Gianluca Pollastri. In silico protein motif discovery and structural analysis. *In Silico Tools for Gene Discovery*, pages 341–353, 2011.

- Vuyani Moses, Rowan Hatherley, and Özlem Tastan Bishop. Bioinformatic characterization of type-specific sequence and structural features in auxiliary activity family 9 proteins. *Biotechnology for Biofuels*, 9(1):1–17, 2016. ISSN 17546834. doi: 10.1186/s13068-016-0655-2.
- Majid Moshirfar, Shaan N Somani, Kathryn M Shmunes, Ladan Espandar, Nikhil S Gokhale, Yasmyne C Ronquillo, and Phillip C Hoopes. A narrative review of microsporidial infections of the cornea. *Ophthalmology and therapy*, 9:265–278, 2020.
- Lamine Mahaman Moustapha, Illiassou Mamane Sadou, Ibrahima Issa Arzika, Laminou Ibrahim Maman, Michel K. Gomgnimbou, Maurice Konkobo, Abdoulaye Diabate, and Etienne Bilgo. First identification of Microsporidia MB in *Anopheles coluzzii* from Zinder City, Niger. *Parasites and Vectors*, 17(1), 12 2024. ISSN 17563305. doi: 10.1186/s13071-023-06059-7.
- Brandon M. Murareanu, Ronesh Sukhdeo, Rui Qu, Jason Jiang, and Aaron W. Reinke. Generation of a microsporidia species attribute database and analysis of the extensive ecological and phenotypic diversity of microsporidia. *bioRxiv*, 2021. URL <https://doi.org/10.1101/2021.02.21.432160>.
- Dilip K Nag and Jed Axelrod. Identification of yeast meiosis-specific genes by differential display. *Methods*, 16(4):423–433, 1998.
- Sirintra Nakjang, Tom A. Williams, Eva Heinz, Andrew K. Watson, Peter G. Foster, Kacper M. Sendra, Sarah E. Heaps, Robert P. Hirt, and T. Martin Embley. Reduction and expansion in microsporidian genome evolution: New insights from comparative genomics. *Genome Biology and Evolution*, 5(12): 2285–2303, 2013. ISSN 17596653. doi: 10.1093/gbe/evt184.
- Godfrey Nattoh, Tracy Maina, Edward E. Makhulu, Lilian Mbaisi, Enock Mararo, Fidel G. Otieno, Tullu Bukhari, Thomas O. Onchuru, Evan Teal, Juan Paredes, Joel L. Bargul, David M. Mburu, Everline A. Onyango, Gabriel Magoma, Steven P. Sinkins, and Jeremy K. Herren. Horizontal Transmission of the Symbiont Microsporidia MB in *Anopheles arabiensis*. *Frontiers in Microbiology*, 12 (July), 2021. ISSN 1664302X. doi: 10.3389/fmicb.2021.647183.

- Godfrey Nattoh, Brenda Onyango, Diana Omoke, Edmond Makhulu, Lilian Mbaisi, Luna Kamau, Maxwell Machani Gesuge, Eric Ochomo, and Jeremy Keith Herren. Microsporidia MB in the primary malaria vector *Anopheles gambiae* s.s. is avirulent and undergoes maternal and horizontal transmission. pages 1–19, 2023. URL <https://doi.org/10.21203/rs.3.rs-2664771/v1>.
- Steve Ndikumana, Adrian Pelin, Alexandre Williot, Justin L. Sanders, Michael Kent, and Nicolas Corradi. Genome Analysis of *Pseudoloma neurophilia*: A Microsporidian Parasite of Zebrafish (*Danio rerio*). *Journal of Eukaryotic Microbiology*, 64(1):18–30, 2017. ISSN 15507408. doi: 10.1111/jeu.12331.
- Ken Nguyen, Xuan Guo, and Yi Pan. Phylogeny in Multiple Sequence Alignments. In *Multiple Biological Sequence Alignment: Scoring Functions, Algorithms and Applications*, chapter 6, pages 103–112. John Wiley & Sons, Inc., 2016. ISBN 9781119273769. doi: <https://doi.org/10.1002/9781119273769.ch6>. URL <https://doi.org/10.1002/9781119273769.ch6>.
- Darryl Nishimura. RepeatMasker. *Biotech Software & Internet Report*, 1:1–4, 2000.
- Dorothy Wavinya Nyamai and Özlem Tastan Bishop. Identification of selective novel hits against plasmodium falciparum prolyl tRNA synthetase active site and a predicted allosteric site using in silico approaches. *International Journal of Molecular Sciences*, 21(11), 2020. ISSN 14220067. doi: 10.3390/ijms21113803.
- Darren J Obbard, Karl H J Gordon, Amy H Buck, and Francis M Jiggins. The evolution of RNAi as a defence against viruses and transposable elements. (October 2008):99–115, 2009. doi: 10.1098/rstb.2008.0168.
- Eric O. Ochomo, Sylvia Milanoi, Bernard Abong’o, Brenda Onyango, Margaret Muchoki, Diana Omoke, Evelyne Olanga, Laban Njoroge, Elijah Juma, and James Dan Otieno. Molecular surveillance leads to the first detection of *Anopheles stephensi* in Kenya. *Research Square*, pages 1–16, 2023.
- Patrick Ojuka, Yap Boum, Lise Denoeud-Ndam, Carolyn Nabasumba, Yolanda Muller, Michael Okia, Juliet Mwanga-Amumpaire, Pierre De Beaudrap, Nat-

- acha Protopopoff, and Jean François Etard. Early biting and insecticide resistance in the malaria vector *Anopheles* might compromise the effectiveness of vector control intervention in Southwestern Uganda. *Malaria Journal*, 2015. ISSN 14752875. doi: 10.1186/s12936-015-0653-z.
- Christian Olsen. Geneious R7: A Bioinformatics Platform for Biologists. *Plant and Animal Genome XXII Conference*, (April):20–21, 2014.
- Umairah Natasya Mohd Omeersheffudin and Suresh Kumar. In silico approach for mining of potential drug targets from hypothetical proteins of bacterial proteome. *International Journal of Molecular Biology*, 4(4):145–152, 2019. doi: 10.15406/ijmboa.2019.04.00111.
- Bugra Ozer, Gizem Gezici, Cem Meydan, and Ugur Sezerman. Multiple sequence alignment based on structural properties. In *2010 5th International Symposium on Health Informatics and Bioinformatics*, pages 39–44, 2010.
- Nitzan Paldi, Eitan Glick, Maayan Oliva, Yaron Zilberberg, Lucie Aubin, Jeffery Pettis, Yanping Chen, and Jay D. Evans. Effective gene silencing in a microsporidian parasite associated with honeybee (*Apis mellifera*) colony declines. *Applied and Environmental Microbiology*, 76(17):5960–5964, 2010. ISSN 00992240. doi: 10.1128/AEM.01067-10.
- Guoqing Pan, Jinshan Xu, Tian Li, Qingyou Xia, Shao Lun Liu, Guojie Zhang, Songgang Li, Chunfeng Li, Handeng Liu, Liu Yang, Tie Liu, Xi Zhang, Zhengli Wu, Wei Fan, Xiaoqun Dang, Heng Xiang, Meilin Tao, Yanhong Li, Junhua Hu, Zhi Li, Lipeng Lin, Jie Luo, Lina Geng, Lin Ling Wang, Mengxian Long, Yongji Wan, Ningjia He, Ze Zhang, Cheng Lu, Patrick J. Keeling, Jun Wang, Zhonghuai Xiang, and Zeyang Zhou. Comparative genomics of parasitic silkworm microsporidia reveal an association between genome expansion and host adaptation. *BMC Genomics*, 14(1), 2013. ISSN 14712164. doi: 10.1186/1471-2164-14-186.
- Guoqing Pan, Jialing Bao, Zhengang Ma, Yue Song, Bing Han, Maoshuang Ran, Chunfeng Li, and Zeyang Zhou. Invertebrate host responses to microsporidia infections. *Developmental and Comparative Immunology*, 83:104–113, 2018. ISSN

18790089. doi: 10.1016/j.dci.2018.02.004. URL <https://doi.org/10.1016/j.dci.2018.02.004>.

Ashutosh K Pandey, Jayashree Pain, Andrew Dancis, and Debkumar Pain. Mitochondria export iron–sulfur and sulfur intermediates to the cytoplasm for iron–sulfur cluster assembly and tRNA thiolation in yeast. *Journal of Biological Chemistry*, 294(24):9489–9502, 2019.

Gianno Pannafino and Eric Alani. Coordinated and independent roles for MLH subunits in DNA repair. *Cells*, 10(4):948, 2021.

Nicolas Parisot, Adrian Pelin, Cyrielle Gasc, Valérie Polonais, Abdel Belkorchia, Johan Panek, Hicham El Alaoui, David G. Biron, Émilie Brassset, Chantal Vaury, Pierre Peyret, Nicolas Corradi, Éric Peyretailade, and Emmanuelle Lerat. Microsporidian genomes harbor a diverse array of transposable elements that demonstrate an ancestry of horizontal exchange with metazoans. *Genome Biology and Evolution*, 6(9):2289–2300, 2014. ISSN 17596653. doi: 10.1093/gbe/evu178.

Eunji Park. Hidden diversity of two intracellular parasites , Microsporidia and Rickettsia, in New Zealand amphipod hosts: patterns and causes. (December): 183 p, 2020. URL <https://ourarchive.otago.ac.nz/handle/10523/10838>.

Eunji Park and Robert Poulin. Revisiting the phylogeny of microsporidia. *International Journal for Parasitology*, 51(XXXX):855–864, 2021. ISSN 18790135. doi: 10.1016/j.ijpara.2021.02.005. URL <https://doi.org/10.1016/j.ijpara.2021.02.005>.

S V Pavlova and S M Zakian. Structural proteins of the SMC (Structural Maintenance of Chromosomes) family and their role in chromatin reorganization. *Russian Journal of Genetics*, 39:1097–1111, 2003.

Katherine Mary Peat. Indications of Sexual Reproduction in Microsporidia Thesis submitted in candidature for the degree of Doctor of Philosophy at Aberystwyth University By. (September), 2011.

- Eric Peyretilade, Hicham El Alaoui, Marie Diogon, Valérie Polonais, Nicolas Parisot, David G. Biron, Pierre Peyret, and Frédéric Delbac. Extreme reduction and compaction of microsporidian genomes. *Research in Microbiology*, 162(6): 598–606, 7 2011. ISSN 09232508. doi: 10.1016/j.resmic.2011.03.004.
- Valérie Polonais, Abdel Belkorchia, Michaël Roussel, Eric Peyretilade, Pierre Peyret, Marie Diogon, and Frédéric Delbac. Identification of two new polar tube proteins related to polar tube protein 2 in the microsporidian *Antonospora locustae*. *FEMS Microbiology Letters*, 2013. ISSN 03781097. doi: 10.1111/1574-6968.12198.
- Jean François Pombert, Karen Luisa Haag, Shadi Beidas, Dieter Ebert, and Patrick J. Keeling. The *Ordospora colligata* genome: Evolution of extreme reduction in microsporidia and host-to-parasite horizontal gene transfer. *mBio*, 6(1), 2015. ISSN 21507511. doi: 10.1128/mBio.02400-14.
- Jean-françois François Pombert, Mohammed Selman, Fabien Burki, Floyd T. Bardell, Laurent Farinelli, Leellen F. Solter, Douglas W. Whitman, Louis M. Weiss, Nicolas Corradi, and Patrick J. Keeling. 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):12638–12643, 2012. ISSN 00278424. doi: 10.1073/pnas.1205020109.
- Joan Pontius, Jean Richelle, and Shoshana J Wodak. Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. *Journal of Molecular Biology*, 264(1):121–136, 1996.
- Simon C. Potter, Aurélien Luciani, Sean R. Eddy, Youngmi Park, Rodrigo Lopez, and Robert D. Finn. HMMER web server: 2018 update. *Nucleic Acids Research*, 46(W1):W200–W204, 2018. ISSN 13624962. doi: 10.1093/nar/gky448.
- R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*, 2017. ISSN 0959-8138. doi: 10.1136/bmj.s4-1.148.879-a. URL <https://www.r-project.org/>.

- T Rajesh and M Jaya. Next-generation sequencing methods. In *Current developments in biotechnology and bioengineering*, pages 143–158. Elsevier, 2017.
- Shahin Ramazi and Javad Zahiri. Post-translational modifications in proteins: resources, tools and prediction methods. *Database*, 2021:baab012, 2021.
- Marilee A Ramesh, Shehre-Banoo Malik, and John M Logsdon. A phylogenomic inventory of meiotic genes: evidence for sex in *Giardia* and an early eukaryotic origin of meiosis. *Current biology*, 15(2):185–191, 2005.
- Bharath Reddy and Richard Fields. Multiple sequence alignment algorithms in bioinformatics. In *Smart Trends in Computing and Communications: Proceedings of SmartCom 2021*, pages 89–98, 2022.
- D. Refardt, E. U. Canning, A. Mathis, S. A. Cheney, N. J. Lafranchi-Tristem, and Dieter Ebert. Small subunit ribosomal DNA phylogeny of microsporidia that infect *Daphnia* (Crustacea: Cladocera). *Parasitology*, 124(4):381–389, 2002. ISSN 00311820. doi: 10.1017/S0031182001001305.
- M A Resnick, J N Kasimos, J C Game, R J Braun, and R M Roth. Changes in DNA during meiosis in a repair-deficient mutant (rad 52) of yeast. *Science*, 212(4494):543–545, 1981.
- Jason M. Rizzo and Michael J. Buck. Key principles and clinical applications of ”next-generation” DNA sequencing, 7 2012. ISSN 19406207.
- Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Andrew J Ballard, Andrew Cowie, Bernardino Romera-paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, and Koray Kavukcuoglu. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(August), 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL <http://dx.doi.org/10.1038/s41586-021-03819-2>.
- Caroline Ross, Caroline Knox, and Özlem Tastan Bishop. Interacting motif networks located in hotspots associated with RNA release are conserved in En-

- terovirus capsids. *FEBS Letters*, 591(12):1687–1701, 2017. ISSN 18733468. doi: 10.1002/1873-3468.12663.
- Guido Van Rossum. Installing Python Modules, 2017. ISSN 18672485.
- Alae Eddine Sahbou, Driss Iraqi, Rachid Mentag, and Slimane Khayi. BuscoPhylo: a webserver for Busco-based phylogenomic analysis for non-specialists. *Scientific Reports*, 12(1):1–5, 2022. ISSN 20452322. doi: 10.1038/s41598-022-22461-0. URL <https://doi.org/10.1038/s41598-022-22461-0>.
- Tanmaya Kumar Sahu, A. R. Rao, Shuchi Vasisht, Nishtha Singh, and U. P. Singh. Computational approaches, databases and tools for in silico motif discovery. *Interdisciplinary Sciences – Computational Life Sciences*, 4(4):239–255, 2012. ISSN 18671462. doi: 10.1007/s12539-012-0141-x.
- Sabine Santucci-Darmanin and Véronique Paquis-Flucklinger. Homologs of MutS and MutL during mammalian meiosis. *Medecine Sciences: M/S*, 19(1):85–91, 2003.
- Sabine Santucci-Darmanin, Sophie Neyton, Françoise Lespinasse, Anne Saunières, Patrick Gaudray, and Véronique Paquis-Flucklinger. The DNA mismatch-repair MLH3 protein interacts with MSH4 in meiotic cells, supporting a role for this MutL homolog in mammalian meiotic recombination. *Human molecular genetics*, 11(15):1697–1706, 2002.
- Eric W. Sayers, Jeffrey Beck, Evan E. Bolton, Devon Bourexis, James R. Brister, Kathi Canese, Donald C. Comeau, Kathryn Funk, Sunghwan Kim, William Klimke, Aron Marchler-Bauer, Melissa Landrum, Stacy Lathrop, Zhiyong Lu, Thomas L. Madden, Nuala O’Leary, Lon Phan, Sanjida H. Rangwala, Valerie A. Schneider, Yuri Skripchenko, Jiyao Wang, Jian Ye, Barton W. Trawick, Kim D. Pruitt, and Stephen T. Sherry. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 49(D1):D10–D17, 2021. ISSN 13624962. doi: 10.1093/nar/gkaa892.
- Hembree S.C., AU Ryan J.R., and AN. Observations on the vertical transmission of a new microsporidian pathogen of *Aedes aegypti* from Thailand [Microbial mosquito control], 1982.

- Erwin Schneider and Sabine Hunke. ATP-binding-cassette (ABC) transport systems: Functional and structural aspects of the ATP-hydrolyzing subunits/domains. *FEMS Microbiology Reviews*, 22(1):1–20, 1998. ISSN 01686445. doi: 10.1016/S0168-6445(98)00002-3.
- Ernst-Jan Scholte, Bart G J Knols, Robert a Samson, and Willem Takken. Entomopathogenic fungi for mosquito control: a review. *Journal of insect science (Online)*, 4:19, 2004. ISSN 1536-2442. doi: 10.1673/031.004.1901. URL <http://jinsectscience.oxfordjournals.org/content/jis/4/1/19.full.pdf>.
- Rahul Semwal, Imlimaong Aier, Utkarsh Raj, and Pritish Kumar Varadwaj. Pr[m]: An Algorithm for Protein Motif Discovery. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):585–592, 2022. ISSN 15579964. doi: 10.1109/TCBB.2020.2999262.
- Igor V. Senderskiy, Sergey A. Timofeev, Elena V. Seliverstova, Olga A. Pavlova, and Viacheslav V. Dolgikh. Secretion of Antonospora (Paranosema) locustae proteins into infected cells suggests an active role of microsporidia in the control of host programs and metabolic processes. *PLoS ONE*, 9(4), 2014. ISSN 19326203. doi: 10.1371/journal.pone.0093585.
- Mohd Shahbaaz, Md Imtaiyaz Hassan, and Faizan Ahmad. Functional annotation of conserved hypothetical proteins from Haemophilus influenzae Rd KW20. *PLoS ONE*, 8(12), 2013. ISSN 19326203. doi: 10.1371/journal.pone.0084263.
- Yanfang Shang, Peng Feng, and Chengshu Wang. Fungi That Infect Insects: Altering Host Behavior and Beyond. *PLoS Pathogens*, 11(8):1–6, 2015. ISSN 15537374. doi: 10.1371/journal.ppat.1005037.
- Qiao Shi, Yi Zhang, Mengqi Ni, Yunhui Bai, Bin Zhou, Jinbin Zheng, and Zhaoxia Cui. Expression characteristics and inhibitory activity of a leucine-rich repeat (lrr)-only protein in the chinese mitten crab, eriocheir sinensis. *Fish & Shellfish Immunology*, 145:109300, 2024.
- Akira Shinohara, Hideyuki Ogawa, and Tomoko Ogawa. Rad51 protein involved in repair and recombination in *S. cerevisiae* is a RecA-like protein. *Cell*, 69(3): 457–470, 1992.

- Shadi Shokralla, Jennifer L. Spall, Joel F. Gibson, and Mehrdad Hajibabaei. Next-generation sequencing technologies for environmental DNA research, 4 2012. ISSN 09621083.
- N. Silaparasetty. Python Programming in Jupyter Notebook. In *Machine Learning Concepts with Python and the Jupyter Notebook Environment*, pages 119–145. Apress, Berkeley, 2020. ISBN 978-1-4842-5967-2. doi: https://doi.org/10.1007/978-1-4842-5967-2{_}7.
- Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, 2015. ISSN 14602059. doi: [10.1093/bioinformatics/btv351](https://doi.org/10.1093/bioinformatics/btv351).
- Selvarajan Sivashankari and Piramanayagam Shanmughavel. Functional annotation of hypothetical proteins – A review. *Bioinformation*, 1(8):335–338, 2006. ISSN 09738894. doi: [10.6026/97320630001335](https://doi.org/10.6026/97320630001335).
- Claudio H. Slamovits, Naomi M. Fast, Joyce S. Law, and Patrick J. Keeling. Genome Compaction and Stability in Microsporidian Intracellular Parasites. *Current Biology*, 14:891–896, 2004. doi: [10.1016/j.cub.2004.07.018](https://doi.org/10.1016/j.cub.2004.07.018).
- Barton E. Slatko, Ashley N. Luck, Stephen L. Dobson, and Jeremy M. Foster. Wolbachia endosymbionts and human disease control, 2014. ISSN 18729428.
- Ryan C. Smith, Joel Vega-Rodríguez, and Marcelo Jacobs-Lorena. The Plasmodium bottleneck: Malaria parasite losses in the mosquito vector. *Memórias do Instituto Oswaldo Cruz*, 109(5):644–661, 2014. ISSN 16788060. doi: [10.1590/0074-0276130597](https://doi.org/10.1590/0074-0276130597).
- Johannes Söding, Andreas Biegert, and Andrei N. Lupas. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, 33(SUPPL. 2):244–248, 2005. ISSN 03051048. doi: [10.1093/nar/gki408](https://doi.org/10.1093/nar/gki408).
- Erik L L Sonnhammer and Anders Krogh. A hidden Markov model for predicting transmembrane helices in protein sequence. *Sixth Int. Conf. on*

Intelligent Systems for Molecular Biology, page 8, 2008. URL papers://4b986d00-906f-493f-a74b-71e29d82b719/Paper/p6291.

Mario Stanke and Burkhard Morgenstern. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research*, 33(SUPPL. 2):465–467, 2005. ISSN 03051048. doi: 10.1093/nar/gki458.

Grant D. Stentiford, David Bass, and Bryony A.P. Williams. Ultimate opportunists—the emergent enterocytozoon group microsporidia. *PLoS Pathogens*, 15(5):1–6, 2019. ISSN 15537374. doi: 10.1371/journal.ppat.1007668.

Alexander V Strunnikov. SMC proteins and chromosome structure. *Trends in Cell Biology*, 8(11):454–459, 1998.

Alexander V Strunnikov and Rolf Jessberger. Structural maintenance of chromosomes (SMC) proteins: conserved molecular properties for multiple biological functions. *European journal of biochemistry*, 263(1):6–13, 1999.

Dmitry Suplatov, Evgeny Kirilin, Vytas Švedas, and A Svendsen. Bioinformatic analysis of protein families to select function-related variable positions. *Understanding enzymes: Function, design, engineering, and analysis*, pages 351–385, 2016.

Dennis Svedberg, Rahel R Winiger, Alexandra Berg, Himanshu Sharma, Christian Tellgren-Roth, Bettina A Debrunner-Vossbrinck, Charles R Vossbrinck, and Jonas Barandun. Functional annotation of a divergent genome using sequence and structure-based similarity. *BMC genomics*, 25(1):6, 2024.

Charles Swanton, Barbara Nicke, and Julian Downward. RNA interference, DNA methylation, and gene silencing: A bright future for cancer therapy? [1]. *Lancet Oncology*, 5(11):653–654, 2004. ISSN 14702045. doi: 10.1016/S1470-2045(04)01604-3.

Damian Szklarczyk, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T. Doncheva, John H. Morris, Peer Bork, Lars J. Jensen, and Christian Von Mering. STRING v11:

- Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 2019. ISSN 13624962. doi: 10.1093/nar/gky1131.
- Damian Szklarczyk, Annika L. Gable, Katerina C. Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T. Doncheva, Marc Legeay, Tao Fang, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612, 2021. ISSN 13624962. doi: 10.1093/nar/gkaa1074.
- Suzannah C. Szumowski and Emily R. Troemel. Microsporidia-Host Interactions. *Current Opinion in Microbiology*, 26(1):10–16, 2015. doi: 10.1016/j.mib.2015.03.006.Microsporidia-Host.
- Hala Tamim El Jarkass and Aaron W. Reinke. The ins and outs of host-microsporidia interactions during invasion, proliferation and exit. *Cellular Microbiology*, 22(11):1–12, 2020. ISSN 14625822. doi: 10.1111/cmi.13247.
- Chuan Yi Tang, Che Lun Hung, Huiru Zheng, Chun Yuan Lin, and Hai Jiang. Novel computational technologies for next-generation sequencing data analysis and their applications, 2015. ISSN 23144378.
- Catherine Texier, Damien Brosseau, Hicham El Alaoui, Guy Méténier, and Christian P. Vivarès. Post-genomics of microsporidia, with emphasis on a model of minimal eukaryotic proteome: A review. *Folia Parasitologica*, 52(1-2):15–22, 2005. ISSN 00155683. doi: 10.14411/fp.2005.003.
- Catherine Texier, Cyril Vidau, Bernard Viguès, Hicham El Alaoui, and Frédéric Delbac. Microsporidia: A model for minimal parasite-host interactions. *Current Opinion in Microbiology*, 13(4):443–449, 2010. ISSN 13695274. doi: 10.1016/j.mib.2010.05.005.
- Chander Jyoti Thakur, Sandeep Saini, Aayushi Notra, and Bhavanshu Chauhan. Deciphering the functional role of hypothetical proteins from *Chloroflexus aurantiacus* J-10-f1 using bioinformatics approach. 9(3):129–139, 2020. doi: 10.22099/mbrc.2020.36894.1495.

- Yuri S Tokarev, Julia M Malysh, Anastasia G. Kononchuk, Elena V. Seliverstova, Andrei N Frolov, and Irma V Issi. Redefinition of *Nosema pyrausta* (Perezia pyraustae Paillot 1927) basing upon ultrastructural and molecular phylogenetic studies. *Parasitology Research*, 114(2):759–761, 2015. ISSN 14321955. doi: 10.1007/s00436-014-4272-3.
- Yuri S. Tokarev, Sergei A. Timofeev, Julia M. Malysh, Alexander A. Tsarev, Anastasia N. Ignatieva, Oksana G. Tomilova, and Vyacheslav V. Dolgikh. Hexokinase as a versatile molecular genetic marker for Microsporidia. *Parasitology*, 146(4):472–478, 4 2019. ISSN 0031-1820. doi: 10.1017/S0031182018001737. URL https://www.cambridge.org/core/product/identifier/S0031182018001737/type/journal_article.
- Emily R. Troemel, Marie Anne Félix, Noah K. Whiteman, Antoine Barrière, and Frederick M. Ausubel. Microsporidia are natural intracellular parasites of the nematode *Caenorhabditis elegans*. *PLoS biology*, 6(12), 2008. ISSN 15457885. doi: 10.1371/journal.pbio.0060309.
- Artur Trzebny, Anna Slodkowicz-Kowalska, James J. Becnel, Neil Sanscrainte, and Mirosława Dabert. A new method of metabarcoding Microsporidia and their hosts reveals high levels of microsporidian infections in mosquitoes (Culicidae). *Molecular Ecology Resources*, 20(6):1486–1504, 2020. ISSN 17550998. doi: 10.1111/1755-0998.13205.
- Artur Trzebny, Justyna Mizera, and Mirosława Dabert. Microsporidians (Microsporidia) parasitic on mosquitoes (Culicidae) in central Europe are often multi-host species. *Journal of Invertebrate Pathology*, 197(August 2022): 107873, 2023. ISSN 10960805. doi: 10.1016/j.jip.2022.107873. URL <https://doi.org/10.1016/j.jip.2022.107873>.
- Anastasios D Tsaousis, Eleni Gentekaki, Laura Eme, Daniel Gaston, and Andrew J Roger. Evolution of the cytosolic iron-sulfur cluster assembly machinery in *Blastocystis* species and other microbial eukaryotes. *Eukaryotic cell*, 13(1): 143–153, 2014.

- Hideo Tsubouchi and G Shirleen Roeder. The Mnd1 protein forms a complex with hop2 to promote homologous chromosome pairing and meiotic double-strand break repair. *Molecular and cellular biology*, 22(9):3078–3088, 2002.
- Eric S Tvedte, Andrew A Forbes, and John M Logsdon Jr. Retention of core meiotic genes across diverse Hymenoptera. *Journal of Heredity*, 108(7):791–806, 2017.
- Albert H Undeen and E Alger. The Effect of the Microsporidan , Nosema on Anopheles stephensi. *Journal of Invertebrate Pathology*, 24:19–24, 1975.
- Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *Biorxiv*, pages 2022–02, 2022.
- Guido Van Rossum and F L Drake. *The python reference manual*. Number March. iUniverse, Indiana, 2000. URL www.python.org.
- Srivats Venkataramanan, Stephen Douglass, Anoop Raj Galivanche, and Tracy Johnson. The chromatin remodeling complex SWI/SNF is a master regulator of meiotic splicing in Saccharomyces cerevisiae. *The FASEB Journal*, 30:802, 2016.
- Anne M Villeneuve and Kenneth J Hillers. Whence meiosis? *Cell*, 106(6):647–650, 2001.
- Christian P. Vivarès and Guy Méténier. The microsporidia genome: Living with minimal genes as an intracellular eukaryote. *Opportunistic Infections: Toxoplasma, Sarcocystis, and Microsporidia*, pages 215–242, 2008. doi: 10.1007/978-1-4020-7846-0{_}12.
- Charles R. Vossbrinck and Bettina A. Debrunner-Vossbrinck. Molecular phylogeny of the Microsporidia: Ecological, ultrastructural and taxonomic considerations. *Folia Parasitologica*, 52(1-2):131–142, 2005. ISSN 00155683. doi: 10.14411/fp.2005.017.

- Charles R Vossbrinck, Theodore G Andreadis, and Louis M Weiss. Phylogenetics: taxonomy and the microsporidia as derived fungi. *Opportunistic infections: toxoplasma, sarcocystis, and microsporidia*, pages 189–213, 2004.
- Ross F. Waller, Carole Jabbour, Nickie C. Chan, Nermin Celik, Vladimir A. Likić, Terrence D. Mulhern, and Trevor Lithgow. Evidence of a reduced and modified mitochondrial protein import apparatus in microsporidian mitosomes. *Eukaryotic Cell*, 8(1):19–26, 2009. ISSN 15359778. doi: 10.1128/EC.00313-08.
- Yan Wang, Hang Zhang, Haolin Zhong, and Zhidong Xue. Protein domain identification methods and online resources, 1 2021. ISSN 20010370.
- Yuan Wang, Jie Chen, Ying Na, Xin-cang Li, Jun-fang Zhou, Wen-hong Fang, and Hong-xin Tan. *Ecytonucleospora hepatopenaei* n. gen. et comb. (Microsporidia: Enterocytozoonidae): A redescription of the Enterocytozoon hepatopenaei (Tourtip et al., 2009), a microsporidian infecting the widely cultivated shrimp *Penaeus vannamei*. *Journal of Invertebrate Pathology*, 201 (August):107988, 2023. ISSN 00222011. doi: 10.1016/j.jip.2023.107988. URL <https://doi.org/10.1016/j.jip.2023.107988>.
- Andrew M. Waterhouse, James B. Procter, David M.A. Martin, Michèle Clamp, and Geoffrey J. Barton. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp033.
- Robert M. Waterhouse, Mathieu Seppey, Felipe A. Simao, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, 35(3):543–548, 2018. ISSN 15371719. doi: 10.1093/molbev/msx319.
- Louis M. Weiss and Aaron W. Reinke. *Microsporidia: Current Advances in Biology*. Springer Nature Switzerland, Cham, Switzerland, 2022. ISBN 9783030933050. doi: <https://doi.org/10.1007/978-3-030-93306-7>.
- Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey

- Kolesnikov, Nathan D Olson, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology*, 37(10):1155–1162, 2019.
- Charles Whittaker, Arran Hamlet, Ellie Sherrard-Smith, Peter Winskill, Gina Cuomo-Dannenburg, Patrick G. T. Walker, Marianne E. Sinka, Samuel Pironon, Ashwani Kumar, Azra C. Ghani, Samir Bhatt, and Thomas S. Churcher. Seasonal dynamics of *Anopheles stephensi* and its implications for mosquito detection and emergent malaria control in the Horn of Africa. *Proceedings of the National Academy of Sciences*, 120(8):1–9, 2023. doi: 10.1073/pnas.22161421201. URL <https://doi.org/10.1073/pnas.22161421201>.
- Miranda M A Whitten, Paul D Facey, Ricardo Del Sol, Meirwyn C Evans, Jacob J Mitchell, Owen G Bodger, Paul J Dyson, R Sol, Mitchell Jj, Bodger Og, Dyson Pj, and Paul J Dyson. Symbiont-mediated RNA interference in insects. *Proc. R. Soc. B*, 283(1825), 2016. ISSN 0962-8452. doi: 10.1098/rspb.2016.0042.
- Ryan R. Wick, Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*, 13(6):1–22, 2017. ISSN 15537358. doi: 10.1371/journal.pcbi.1005595.
- Markus Wiederstein and Manfred J. Sippl. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research*, 35(SUPPL.2):407–410, 2007. ISSN 03051048. doi: 10.1093/nar/gkm290.
- Bryony A P Williams, Robert P Hirt, and John M Lucocq. A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*. 418(August):865–869, 2002. doi: 10.1038/nature00990.1.
- Bryony A.P. Williams. Unique physiology of host-parasite interactions in microsporidia infections. *Cellular Microbiology*, 11(11):1551–1560, 2009. ISSN 14625814. doi: 10.1111/j.1462-5822.2009.01362.x.
- Bryony A.P. Williams, Renny C.H. Lee, James J. Becnel, Louis M. Weiss,

- Naomi M. Fast, and Patrick J. Keeling. Genome sequence surveys of *Brachyola algerae* and *Edhazardia aedis* reveal microsporidia with low gene densities. *BMC Genomics*, 9, 4 2008. ISSN 14712164. doi: 10.1186/1471-2164-9-200.
- Marni Williams, Brady J. Summers, and Richard H.G. Baxter. Biophysical Analysis of *Anopheles gambiae* Leucine-Rich Repeat Proteins APL1A1, APLB and APL1C and Their Interaction with LRIM1. *PLoS ONE*, 10(3):1–20, 2015. ISSN 19326203. doi: 10.1371/journal.pone.0118911.
- Alexandra R Willis and Aaron W Reinke. Factors That determine Microsporidia infection and host specificity. In *Microsporidia: Current Advances in Biology*, pages 91–114. Springer, 2022.
- Dominic Wiredu Boakye, Pattana Jaroenlak, Anuphap Prachumwat, Tom A. Williams, Kelly S. Bateman, Ornthuma Itsathitphaisarn, Kallaya Sritunyaluck-sana, Konrad H. Paszkiewicz, Karen A. Moore, Grant D. Stentiford, and Bryony A.P. Williams. Decay of the glycolytic pathway and adaptation to intranuclear parasitism within Enterocytozoonidae microsporidia. *Environmental Microbiology*, 19(5):2077–2089, 2017. ISSN 14622920. doi: 10.1111/1462-2920.13734.
- Murray Wittner and Louis M. Weiss. *The Microsporidia and Microsporidiosis*. 1999.
- Seon-Kyung Woo, Chang-Beom Park, and Seong-Whan Lee. Protein secondary structure prediction using sequence profile and conserved domain profile. In *International Conference on Intelligent Computing*, pages 1–10, 2005.
- Derrick E. Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1):1–13, 2019. ISSN 1474760X. doi: 10.1186/s13059-019-1891-0.
- World Health Organization. *World malaria report 2022*. 2022. ISBN 9789240064898.
- Kevin M Wright, Brian Arnold, Katherine Xue, Maria Šurinová, Jeremy O’Connell, and Kirsten Bomblies. Selection on meiosis genes in diploid and

- tetraploid *Arabidopsis arenosa*. *Molecular biology and evolution*, 32(4):944–955, 2015.
- Shannon E. Wright and Peter K. Todd. Native functions of short tandem repeats. *eLife*, 12, 3 2023. ISSN 2050084X. doi: 10.7554/elife.84043.
- Cathy H. Wu, Anastasia Nikolskaya, Hongzhan Huang, Lai Su Yeh, Darren A. Natale, C. R. Vinayaka, Zhang Zhi Hu, Raja Mazumder, Sandeep Kumar, Panagiotis Kourtesis, Robert S. Ledley, Baris E. Suzek, Leslie Arminski, Yongxing Chen, Jian Zhang, Jorge Louie Cardenas, Sehee Chung, Jorge Castro-Alvear, Georgi Dinkov, and Winona C. Barker. PIRSF: Family classification system at the Protein Information Resource. *Nucleic Acids Research*, 32(DATABASE ISS.):112–114, 2004. ISSN 03051048. doi: 10.1093/nar/gkh097.
- Zhexin Xiang. Advances in homology protein structure modeling. *Current Protein and Peptide Science*, 7(3):217–227, 2006.
- Chao Xu and Jinrong Min. Structure and function of WD40 domain proteins. *Protein and Cell*, 2(3):202–214, 2011. ISSN 16748018. doi: 10.1007/s13238-011-1018-1.
- Jinshan Xu, Guoqing Pan, Lin Fang, Jun Li, Xiangjun Tian, Tian Li, Zeyang Zhou, and Zhonghuai Xiang. The varying microsporidian genome: Existence of long-terminal repeat retrotransposon in domesticated silkworm parasite *Nosema bombycis*. *International Journal for Parasitology*, 36(9):1049–1056, 2006. ISSN 00207519. doi: 10.1016/j.ijpara.2006.04.010.
- Shinpei Yamaguchi, Kwonho Hong, Rui Liu, Li Shen, Azusa Inoue, Dinh Diep, Kun Zhang, and Yi Zhang. Tet1 controls meiosis by regulating meiotic gene expression. *Nature*, 492(7429):443–447, 2012.
- Donglin Yang, Guoqing Pan, Xiaoqun Dang, Yawei Shi, Chunfeng Li, Pai Peng, Bo Luo, Maofoi Bian, Yue Song, Cheng Ma, Jie Chen, Zhengang Ma, Lina Geng, Zhi Li, Rui Tian, Cuifang Wei, and Zeyang Zhou. Interaction and assembly of two novel proteins in the spore wall of the microsporidian species *Nosema bombycis* and their roles in adherence to and infection of host cells. *Infection and Immunity*, 83(4):1715–1731, 2015. ISSN 10985522. doi: 10.1128/IAI.03155-14.

- Donglin Yang, Lixia Pan, Zhongzhu Chen, Huihui Du, Bo Luo, Jie Luo, and Guoqing Pan. The roles of microsporidia spore wall proteins in the spore wall formation and polar tube anchorage to spore wall during development and infection processes. *Experimental parasitology*, 187:93–100, 2018.
- Hongxing Yang, Pingli Lu, Yingxiang Wang, and Hong Ma. The transcriptome landscape of Arabidopsis male meiocytes from high-throughput sequencing: the complexity and evolution of the meiotic process. *The Plant Journal*, 65(4):503–516, 2011.
- Mingzhang Yang, Myra K. Derbyshire, Roxanne A. Yamashita, and Aron Marchler-Bauer. NCBI’s Conserved Domain Database and Tools for Protein Domain Analysis. *Current Protocols in Bioinformatics*, 69(1), 3 2020. ISSN 1934340X. doi: 10.1002/cpbi.90.
- K Yokomori. SMC protein complexes and the maintenance of chromosome integrity. *Protein Complexes that Modify Chromatin*, pages 79–112, 2003.
- Shuguang Yuan, H. C. Stephen Chan, and Zhenquan Hu. Using PyMOL as a platform for computational drug design. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 7(2):1–10, 2017. ISSN 17590884. doi: 10.1002/wcms.1298.
- Feng Yue, Jian Shi, and Jijun Tang. Simultaneous phylogeny reconstruction and multiple sequence alignment. *BMC bioinformatics*, 10(1):1–9, 2009.
- Baiyu Zhang, Zhichao Zhang, Saijiang Yong, Shuyang Yu, Hui Feng, Maozhu Yin, Wenwu Ye, Yuanchao Wang, and Min Qiu. An oomycete-specific leucine-rich repeat-containing protein is involved in zoospore flagellum development in phytophthora sojae. *Phytopathology®*, 112(11):2351–2359, 2022.
- Wei Zhang and Tianwen Wang. Bioinformatics-aided Protein Sequence Analysis and Engineering. *Current Protein and Peptide Science*, 24(6):477–487, 2023.
- Liangbiao Zheng. Molecular approaches to mosquito parasite interactions. *Archives of Insect Biochemistry and Physiology*, 34, 1 1997. doi: 10.1002/(sici)1520-6327(1997)34:

1<1::aid-arch1>3.3.co;2-r. URL <https://discovery.researcher.life/article/molecular-approaches-to-mosquito-parasite-interactions/e18b391c528f328db8dae17657936311>.

Ning Zheng, Ernest Fraenkel, Carl O. Pabo, and Nikola P. Pavletich. Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. *Genes and Development*, 13(6):666–674, 1999. ISSN 08909369. doi: 10.1101/gad.13.6.666.

Lukas Zimmermann, Andrew Stephens, Seung Zin Nam, David Rau, Jonas Kübler, Marko Lozajic, Felix Gabler, Johannes Söding, Andrei N. Lupas, and Vikram Alva. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology*, 430(15):2237–2243, 7 2018. ISSN 10898638. doi: 10.1016/j.jmb.2017.12.007. URL <https://doi.org/10.1016/j.jmb.2017.12.007>.

Peter Zwickl, Anja Grziwa, Gabriela Piühler, Burkhardt Dahlmann, Friedrich Lottspeich, and Wolfgang Baumeister. Primary Structure of the Thermoplasma Proteasome and Its Implications for the Structure , Function , and Evolution of the Multicatalytic Proteinase7. pages 964–972, 1992.

Appendices

Appendix 1: Subcellular localization prediction of *V. corneae* HPs using different online tools. The hydrophobic proteins also had membrane-spanning regions.

Table A1: Subcellular localization prediction of HPs using different online tools. The hydrophobic proteins also had membrane-spanning regions.

Sequence ID	GRAVY SCORE	Grand Average of Hydropathy (GRAVY)	Membrane prediction using TargetP	Membrane prediction using SignalP	Membrane prediction using DeepTMHMM	Membrane region prediction using TMHMM	TMHMM: Y/N indicator if a transmembrane region was predicted or not	Phobius: Y/N indicator if a signal peptide was predicted or not (SP)
VICG_02177	Hydrophilic	-0.69676	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_02113	Hydrophilic	-0.11802	OTHER	OTHER	GLOB	Topology=o	NULL	Y
VICG_02001	Hydrophilic	-0.70339	OTHER	OTHER	GLOB	Topology=i	NULL	NULL
VICG_02006	Hydrophilic	-0.88433	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00267	Hydrophilic	-1.06728	OTHER	OTHER	GLOB	Topology=o	NULL	NULL

Continued on next page

Table A1: Subcellular localization prediction of HPs using different online tools. The hydrophobic proteins also had membrane-spanning regions. (Continued)

Sequence ID	GRAVY SCORE	Grand Average of Hydropathy (GRAVY)	Membrane prediction using TargetP	Membrane prediction using SignalP	Membrane prediction using DeepTMHMM	Membrane region prediction using TMHMM	TMHMM: Y/N indicator if a transmembrane region was predicted or not	Phobius: Y/N indicator if a signal peptide was predicted or not (SP)
VICG_01972	Hydrophilic	-0.39593	OTHER	OTHER	GLOB	Topology=i	NULL	NULL
VICG_01905	Hydrophilic	-0.87866	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_02179	Hydrophobic	0.082759	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01970	Hydrophobic	0.688824	OTHER	OTHER	TM	Topology=i7-29o44- 66i87-109o122-144i	Y	NULL
VICG_02174	Hydrophobic	0.114655	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01935	Hydrophilic	-0.32743	OTHER	OTHER	SP	Topology=o	NULL	Y
VICG_01923	Hydrophobic	0.061765	SP	SP(Sec/SPI)	SP+TM	Topology=o106- 128i	Y	Y
VICG_02133	Hydrophilic	-0.25259	SP	OTHER	SP	Topology=i12-31o	Y	NULL
VICG_01861	Hydrophilic	-1.06019	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01938	Hydrophilic	-1.06094	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_02124	Hydrophilic	-0.04245	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01869	Hydrophilic	-0.81527	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01837	Hydrophilic	-0.14605	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01777	Hydrophilic	-1.00604	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01700	Hydrophilic	-0.14393	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01710	Hydrophobic	0.674902	OTHER	OTHER	TM	Topology=i50- 72o82-101i122- 144o164-186i199- 221o	Y	NULL
VICG_01745	Hydrophobic	0.240476	OTHER	OTHER	GLOB	Topology=o213- 235i	Y	NULL
VICG_00267	Hydrophilic	-1.06728	OTHER	OTHER	GLOB	Topology=o	NULL	NULL

Continued on next page

Table A1: Subcellular localization prediction of HPs using different online tools. The hydrophobic proteins also had membrane-spanning regions. (Continued)

Sequence ID	GRAVY SCORE	Grand Average of Hydropathy (GRAVY)	Membrane prediction using TargetP	Membrane prediction using SignalP	Membrane prediction using DeepTMHMM	Membrane region prediction using TMHMM	TMHMM: Y/N indicator if a transmembrane region was predicted or not	Phobius: Y/N indicator if a signal peptide was predicted or not (SP)
VICG_01980	Hydrophilic	-0.54874	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_02005	Hydrophilic	-0.69649	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01689	Hydrophilic	-0.95541	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01887	Hydrophobic	0.566839	OTHER	OTHER	TM	Topology=i13-35o50-72i79-101o170-192i	Y	NULL
VICG_01597	Hydrophilic	-0.21944	SP	OTHER	SP	Topology=i7-29o	Y	NULL
VICG_02135	Hydrophilic	-0.45833	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01848	Hydrophilic	-0.30722	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01927	Hydrophilic	-0.12158	OTHER	OTHER	GLOB	Topology=o	NULL	Y
VICG_01768	Hydrophilic	-0.59106	SP	SP(Sec/SPI)	SP	Topology=o	NULL	Y
VICG_01808	Hydrophobic	0.481992	OTHER	OTHER	TM	Topology=o104-126i135-157o181-203i230-252o	Y	NULL
VICG_01556	Hydrophobic	0.065373	SP	SP(Sec/SPI)	SP	Topology=o	NULL	Y
VICG_01738	Hydrophobic	0.648669	OTHER	OTHER	SP+TM	Topology=o63-85i104-126o136-155i172-194o209-231i	Y	Y
VICG_01465	Hydrophilic	-0.79352	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01987	Hydrophilic	-1.05979	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01677	Hydrophilic	-0.24596	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01754	Hydrophilic	-0.55	OTHER	OTHER	TM	Topology=i13-35o	Y	Y
VICG_00267	Hydrophilic	-1.06728	OTHER	OTHER	GLOB	Topology=o	NULL	NULL

Continued on next page

Table A1: Subcellular localization prediction of HPs using different online tools. The hydrophobic proteins also had membrane-spanning regions. (Continued)

Sequence ID	GRAVY SCORE	Grand Average of Hydropathy (GRAVY)	Membrane prediction using TargetP	Membrane prediction using SignalP	Membrane prediction using DeepTMHMM	Membrane region prediction using TMHMM	TMHMM: Y/N indicator if a transmembrane region was predicted or not	Phobius: Y/N indicator if a signal peptide was predicted or not (SP)
VICG_01951	Hydrophilic	-0.16023	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01799	Hydrophilic	-0.24781	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01914	Hydrophilic	-0.74837	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01509	Hydrophilic	-0.61302	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01658	Hydrophilic	-0.73554	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01274	Hydrophilic	-0.99966	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01693	Hydrophobic	0.515054	OTHER	OTHER	TM	Topology=o40- 62i83-105o159-181i	Y	NULL
VICG_01289	Hydrophilic	-0.27742	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01212	Hydrophilic	-0.35167	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01304	Hydrophilic	-0.02918	OTHER	OTHER	TM	Topology=i17-39o	Y	NULL
VICG_01900	Hydrophilic	-0.07978	OTHER	OTHER	GLOB	Topology=o15-33i	Y	NULL
VICG_01319	Hydrophilic	-0.26066	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01816	Hydrophobic	0.583077	OTHER	OTHER	TM	Topology=i65- 87o107-129i	Y	NULL
VICG_01200	Hydrophobic	0.20625	OTHER	OTHER	GLOB	Topology=i21-42o	Y	NULL
VICG_01237	Hydrophilic	-0.62159	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01826	Hydrophilic	-0.711	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01611	Hydrophilic	-0.13504	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01756	Hydrophilic	-0.96766	OTHER	OTHER	TM	Topology=o11-30i	Y	Y
VICG_00267	Hydrophilic	-1.06728	OTHER	OTHER	GLOB	Topology=o	NULL	NULL

Continued on next page

Table A1: Subcellular localization prediction of HPs using different online tools. The hydrophobic proteins also had membrane-spanning regions. (Continued)

Sequence ID	GRAVY SCORE	Grand Average of Hydropathy (GRAVY)	Membrane prediction using TargetP	Membrane prediction using SignalP	Membrane prediction using DeepTMHMM	Membrane region prediction using TMHMM	TMHMM: Y/N indicator if a transmembrane region was predicted or not	Phobius: Y/N indicator if a signal peptide was predicted or not (SP)
VICG_01408	Hydrophobic	0.502076	OTHER	OTHER	TM	Topology=i100- 122o132-151i186- 208o223-245i258- 280o	Y	NULL
VICG_01180	Hydrophobic	0.458904	OTHER	OTHER	TM	Topology=i12- 34o49-71i92- 114o147-169i	Y	NULL
VICG_01153	Hydrophilic	-0.68	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01741	Hydrophobic	0.539643	OTHER	OTHER	SP+TM	Topology=i12- 34o72-94i115- 137o152-171i184- 206o226-248i	Y	NULL
VICG_01384	Hydrophilic	-0.93077	OTHER	OTHER	GLOB	Topology=i	NULL	NULL
VICG_01567	Hydrophilic	-0.05888	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01282	Hydrophilic	-0.10743	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01399	Hydrophilic	-0.30317	OTHER	OTHER	GLOB	Topology=o	NULL	Y
VICG_01307	Hydrophilic	-0.73058	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01322	Hydrophilic	-0.49864	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00975	Hydrophilic	-0.08995	OTHER	OTHER	SP	Topology=o	NULL	Y
VICG_01613	Hydrophilic	-0.188	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00985	Hydrophilic	-0.09758	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01502	Hydrophilic	-0.37422	OTHER	OTHER	SP	Topology=o	NULL	NULL
VICG_01507	Hydrophilic	-1.01338	SP	SP(Sec/SPI)	SP	Topology=o	NULL	Y
VICG_00267	Hydrophilic	-1.06728	OTHER	OTHER	GLOB	Topology=o	NULL	NULL

Continued on next page

Table A1: Subcellular localization prediction of HPs using different online tools. The hydrophobic proteins also had membrane-spanning regions. (Continued)

Sequence ID	GRAVY SCORE	Grand Average of Hydropathy (GRAVY)	Membrane prediction using TargetP	Membrane prediction using SignalP	Membrane prediction using DeepTMHMM	Membrane region prediction using TMHMM	TMHMM: Y/N indicator if a transmembrane region was predicted or not	Phobius: Y/N indicator if a signal peptide was predicted or not (SP)
VICG_01243	Hydrophilic	-0.25938	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01248	Hydrophobic	0.082308	SP	SP(Sec/SPI)	GLOB	Topology=o	NULL	Y
VICG_01030	Hydrophilic	-0.51852	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00875	Hydrophilic	-0.49032	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01363	Hydrophilic	-0.44333	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00885	Hydrophilic	-0.6404	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00922	Hydrophilic	-0.93396	OTHER	OTHER	GLOB	Topology=i	NULL	NULL
VICG_01461	Hydrophilic	-0.31196	OTHER	OTHER	SP	Topology=o	NULL	Y
VICG_01186	Hydrophilic	-0.63333	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01481	Hydrophilic	-0.27984	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01323	Hydrophilic	-0.29854	OTHER	OTHER	TM	Topology=o166-188i	Y	NULL
VICG_00937	Hydrophilic	-0.263	OTHER	OTHER	TM	Topology=o67-89i	Y	NULL
VICG_01204	Hydrophilic	-0.40321	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01229	Hydrophobic	0.178713	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00764	Hydrophobic	0.042508	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00925	Hydrophilic	-0.19878	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00956	Hydrophilic	-0.95016	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00767	Hydrophilic	-0.04833	OTHER	OTHER	TM	Topology=i21-40o	Y	NULL
VICG_01197	Hydrophilic	-0.03108	SP	SP(Sec/SPI)	GLOB	Topology=o	NULL	Y
VICG_00630	Hydrophilic	-0.5927	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00267	Hydrophilic	-1.06728	OTHER	OTHER	GLOB	Topology=o	NULL	NULL

Continued on next page

Table A1: Subcellular localization prediction of HPs using different online tools. The hydrophobic proteins also had membrane-spanning regions. (Continued)

Sequence ID	GRAVY SCORE	Grand Average of Hydropathy (GRAVY)	Membrane prediction using TargetP	Membrane prediction using SignalP	Membrane prediction using DeepTMHMM	Membrane region prediction using TMHMM	TMHMM: Y/N indicator if a transmembrane region was predicted or not	Phobius: Y/N indicator if a signal peptide was predicted or not (SP)
VICG_00876	Hydrophilic	-0.81125	OTHER	OTHER	GLOB	Topology=i	NULL	NULL
VICG_00671	Hydrophobic	0.57619	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00686	Hydrophobic	0.373671	OTHER	OTHER	GLOB	Topology=o10- 32i63-85o95- 117i136-158o173- 195i	Y	NULL
VICG_00456	Hydrophobic	0.673799	OTHER	OTHER	TM	Topology=i84- 106o121-143i150- 172o	Y	NULL
VICG_00461	Hydrophilic	-0.97283	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00696	Hydrophilic	-0.43852	SP	SP(Sec/SPI)	SP	Topology=o10-27i	Y	Y
VICG_00982	Hydrophilic	-0.86299	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00987	Hydrophilic	-0.42963	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00711	Hydrophilic	-0.63924	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01154	Hydrophilic	-0.9	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00726	Hydrophilic	-0.44824	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00492	Hydrophilic	-0.659	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01017	Hydrophilic	-1.02007	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00887	Hydrophilic	-0.8613	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_01071	Hydrophilic	-0.55788	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00505	Hydrophilic	-0.52288	OTHER	OTHER	GLOB	Topology=i	NULL	NULL
VICG_00832	Hydrophilic	-0.14173	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00924	Hydrophilic	-0.28768	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00267	Hydrophilic	-1.06728	OTHER	OTHER	GLOB	Topology=o	NULL	NULL

Continued on next page

Table A1: Subcellular localization prediction of HPs using different online tools. The hydrophobic proteins also had membrane-spanning regions. (Continued)

Sequence ID	GRAVY SCORE	Grand Average of Hydropathy (GRAVY)	Membrane prediction using TargetP	Membrane prediction using SignalP	Membrane prediction using DeepTMHMM	Membrane region prediction using TMHMM	TMHMM: Y/N indicator if a transmembrane region was predicted or not	Phobius: Y/N indicator if a signal peptide was predicted or not (SP)
VICG_00550	Hydrophilic	-0.68673	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00969	Hydrophilic	-0.3242	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00984	Hydrophilic	-0.50437	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00674	Hydrophobic	0.036254	SP	SP(Sec/SPI)	SP	Topology=o	NULL	Y
VICG_00783	Hydrophilic	-0.46058	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00798	Hydrophilic	-0.17024	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00724	Hydrophilic	-0.63729	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00874	Hydrophilic	-1.02951	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00328	Hydrophilic	-0.51106	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00385	Hydrophilic	-0.54934	SP	SP(Sec/SPI)	GLOB	Topology=o	NULL	Y
VICG_00410	Hydrophobic	0.90708	OTHER	OTHER	TM	Topology=o31- 53i66-88o117- 139i159-181o196- 218i	Y	NULL
VICG_00513	Hydrophilic	-1.11364	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00008	Hydrophilic	-0.25846	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00730	Hydrophilic	-0.27687	OTHER	OTHER	TM	Topology=i81- 103o118-140i	Y	NULL
VICG_00558	Hydrophilic	-0.09459	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00563	Hydrophilic	-0.81575	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00790	Hydrophobic	0.565789	OTHER	OTHER	TM	Topology=i9-31o41- 61i82-104o119- 141i154-176o	Y	NULL
VICG_00267	Hydrophilic	-1.06728	OTHER	OTHER	GLOB	Topology=o	NULL	NULL

Continued on next page

Table A1: Subcellular localization prediction of HPs using different online tools. The hydrophobic proteins also had membrane-spanning regions. (Continued)

Sequence ID	GRAVY SCORE	Grand Average of Hydropathy (GRAVY)	Membrane prediction using TargetP	Membrane prediction using SignalP	Membrane prediction using DeepTMHMM	Membrane region prediction using TMHMM	TMHMM: Y/N indicator if a transmembrane region was predicted or not	Phobius: Y/N indicator if a signal peptide was predicted or not (SP)
VICG_00830	Hydrophilic	-0.58931	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00603	Hydrophilic	-0.36027	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00103	Hydrophilic	-0.24812	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00845	Hydrophilic	-0.88365	OTHER	OTHER	TM	Topology=o	NULL	NULL
VICG_00529	Hydrophilic	-0.12813	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00631	Hydrophobic	0.364205	OTHER	OTHER	GLOB	Topology=o36- 58i118-140o155- 177i189-211o	Y	NULL
VICG_00656	Hydrophobic	0.017411	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00101	Hydrophilic	-0.00758	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00316	Hydrophilic	-0.99076	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00672	Hydrophilic	-0.105	OTHER	OTHER	GLOB	Topology=i	NULL	NULL
VICG_00574	Hydrophilic	-0.87179	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00326	Hydrophilic	-0.82799	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00136	Hydrophilic	-0.725	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00702	Hydrophilic	-0.93121	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00223	Hydrophilic	-0.76561	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00366	Hydrophilic	-0.57225	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00282	Hydrophobic	0.015966	OTHER	OTHER	SP	Topology=i20-42o	Y	NULL
VICG_00297	Hydrophilic	-0.74894	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00201	Hydrophilic	-0.96047	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00391	Hydrophilic	-0.26957	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00267	Hydrophilic	-1.06728	OTHER	OTHER	GLOB	Topology=o	NULL	NULL

Continued on next page

Table A1: Subcellular localization prediction of HPs using different online tools. The hydrophobic proteins also had membrane-spanning regions. (Continued)

Sequence ID	GRAVY SCORE	Grand Average of Hydropathy (GRAVY)	Membrane prediction using TargetP	Membrane prediction using SignalP	Membrane prediction using DeepTMHMM	Membrane region prediction using TMHMM	TMHMM: Y/N indicator if a transmembrane region was predicted or not	Phobius: Y/N indicator if a signal peptide was predicted or not (SP)
VICG_00268	Hydrophobic	0.070879	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00327	Hydrophilic	-0.99679	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00019	Hydrophilic	-0.36092	SP	SP(Sec/SPI)	SP	Topology=o	NULL	Y
VICG_00024	Hydrophilic	-0.07576	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00044	Hydrophilic	-0.58009	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00074	Hydrophilic	-0.88452	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00387	Hydrophilic	-0.44378	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00020	Hydrophilic	-0.43085	SP	SP(Sec/SPI)	SP	Topology=o	NULL	Y
VICG_00329	Hydrophilic	-0.49825	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00100	Hydrophilic	-0.17608	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00364	Hydrophilic	-1.0297	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00165	Hydrophilic	-0.35819	OTHER	OTHER	TM	Topology=i288-310o314-336i	Y	NULL
VICG_00429	Hydrophilic	-0.11149	SP	OTHER	SP	Topology=o	NULL	Y
VICG_00225	Hydrophilic	-1.235	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00052	Hydrophilic	-0.92065	OTHER	OTHER	GLOB	Topology=i	NULL	NULL
VICG_00057	Hydrophilic	-0.81976	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00162	Hydrophilic	-0.69771	OTHER	OTHER	GLOB	Topology=o	NULL	NULL
VICG_00182	Hydrophobic	0.144531	SP	SP(Sec/SPI)	SP+TM	Topology=o93-115i	Y	Y
VICG_00267	Hydrophilic	-1.06728	OTHER	OTHER	GLOB	Topology=o	NULL	NULL

^a Significant scores are highlighted in the output by an asterisk.

GLOB - Globular; SP - Signal Peptide; Y - Yes (Present; N - No (Absent); OTHER - No signal peptide found; Sec/SPI - Signal peptide protein transported by the Sec translocon and cleaved by Signal Peptidase I ; SP+TM - Signal peptide with a transmembrane region ; "Topology=o" - outside the transmembrane region; "Topology = i" - inside the transmembrane region; "Topology=o106-128i " - Amino acid position 0-106

lies outside, 106-128 is the transmembrane domain and 128-end of the protein is inside the region.

Appendix 2: Overview of the KEGG pathway and GO analysis of *V. corneae* HPs

Table A2: Overview of the KEGG pathway and GO analysis of *V. corneae* HPs

# Gene name	KO ID	Threshold	Score	E-value	KO definition
* XP_007603502.1hypotheticalproteinVICG_00049	K07565	39.47	195.2	9.70E-60	60S ribosome subunit biogenesis protein NIP7
XP_007603503.1hypotheticalproteinVICG_00050	K02900	102.7	167.3	3.70E-51	large subunit ribosomal protein L27Ae
* XP_007603534.1hypotheticalproteinVICG_00081	K14831	103.43	223.2	5.90E-68	protein MAK16
* XP_007603549.1hypotheticalproteinVICG_00096	K15264	231.47	294.3	1.70E-89	25S rRNA (cytosine2278-C5)-methyltransferase [EC:2.1.1.311]
* XP_007603555.1hypotheticalproteinVICG_00102	K02324	395.57	1437.4	0	DNA polymerase epsilon subunit 1 [EC:2.7.7.7]
* XP_007603586.1hypotheticalproteinVICG_00133	K09467	102.03	124.1	1.30E-37	zinc-finger protein CreA/MIG
* XP_007603650.1hypotheticalproteinVICG_00197	K03136	42.67	132.7	1.80E-40	transcription initiation factor TFIIE subunit alpha
* XP_007603684.1hypotheticalproteinVICG_00231	K11884	123.77	231.6	8.80E-71	RNA-binding protein PNO1
* XP_007603688.1hypotheticalproteinVICG_00235	K14006	287.87	557.8	5.40E-169	protein transport protein SEC23
* XP_007603795.1hypotheticalproteinVICG_00342	K11699	136.37	570.2	1.00E-172	RNA-dependent RNA polymerase [EC:2.7.7.48]
* XP_007603972.1hypotheticalproteinVICG_00520	K10991	48.5	49.7	3.30E-15	DNA repair protein Swi5/Sae3
* XP_007605535.1hypotheticalproteinVICG_02090	K19613	320.8	374.1	8.60E-114	leucine-rich repeat protein SHOC2

Continued on next page

Table A2: Overview of the KEGG pathway and GO analysis of *V. corneae* HPs (Continued)

# Gene name	KO ID	Threshold	Score	E-value	KO definition
* XP_007603976.1hypotheticalproteinVICG_00524	K11253	217.8	231.3	2.50E-70	histone H3
* XP_007603998.1hypotheticalproteinVICG_00546	K03038	194	265.2	9.20E-81	26S proteasome regulatory subunit N8
* XP_007604025.1hypotheticalproteinVICG_00573	K06961	67.1	238	1.90E-72	ribosomal RNA assembly protein
* XP_007604027.1hypotheticalproteinVICG_00575	K03142	143.47	217.8	2.10E-66	transcription initiation factor TFIIH subunit 2
* XP_007604045.1hypotheticalproteinVICG_00593	K00981	53.8	184.5	2.80E-56	phosphatidate cytidyltransferase [EC:2.7.7.41]
* XP_007604066.1hypotheticalproteinVICG_00614	K19613	320.8	322.5	3.80E-98	leucine-rich repeat protein SHOC2
* XP_007604198.1hypotheticalproteinVICG_00747	K11593	439.33	645.6	2.00E-195	eukaryotic translation initiation factor 2C
* XP_007604216.1hypotheticalproteinVICG_00765	K02974	30.53	55.9	3.70E-17	small subunit ribosomal protein S24e
* XP_007604222.1hypotheticalproteinVICG_00771	K02321	119.73	179.4	8.20E-55	DNA polymerase alpha subunit B
* XP_007604260.1hypotheticalproteinVICG_00811	K02912	69.33	150.2	7.50E-46	large subunit ribosomal protein L32e
* XP_007604292.1hypotheticalproteinVICG_00843	K01174	70.97	110.9	6.30E-34	micrococcal nuclease [EC:3.1.31.1]
* XP_007604335.1hypotheticalproteinVICG_00886	K12580	116.73	138	4.40E-42	CCR4-NOT transcription complex subunit 3
* XP_007604339.1hypotheticalproteinVICG_00890	K06072	94.2	260.3	2.20E-79	deoxyhypusine monooxygenase [EC:1.14.99.29]
* XP_007605535.1hypotheticalproteinVICG_02090	K19613	320.8	374.1	8.60E-114	leucine-rich repeat protein SHOC2

Continued on next page

Table A2: Overview of the KEGG pathway and GO analysis of *V. corneae* HPs (Continued)

# Gene name	KO ID	Threshold	Score	E-value	KO definition
* XP_007604344.1hypotheticalproteinVICG_00897	K14852	38.47	77.7	9.50E-24	regulator of ribosome biosynthesis
* XP_007604350.1hypotheticalproteinVICG_00903	K01262	246.6	292.9	4.80E-89	Xaa-Pro aminopeptidase [EC:3.4.11.9]
* XP_007604487.1hypotheticalproteinVICG_01041	K14834	108.77	135.5	2.10E-41	nucleolar complex protein 3
* XP_007604512.1hypotheticalproteinVICG_01066	K00600	79.5	663.7	3.30E-201	glycine hydroxymethyltransferase [EC:2.1.2.1]
* XP_007604572.1hypotheticalproteinVICG_01126	K00698	150.03	423.3	2.10E-128	chitin synthase [EC:2.4.1.16]
* XP_007604630.1hypotheticalproteinVICG_01184	K00287	53.83	87.1	1.20E-26	dihydrofolate reductase [EC:1.5.1.3]
* XP_007604760.1hypotheticalproteinVICG_01314	K02925	332.17	587.5	4.00E-178	large subunit ribosomal protein L3e
* XP_007604760.1hypotheticalproteinVICG_01314	K02906	115.77	116.7	1.20E-35	large subunit ribosomal protein L3
* XP_007604773.1hypotheticalproteinVICG_01327	K03357	94.9	95.6	3.30E-29	anaphase-promoting complex subunit 10
* XP_007604813.1hypotheticalproteinVICG_01367	K14007	221.73	355.7	4.10E-108	protein transport protein SEC24
* XP_007604824.1hypotheticalproteinVICG_01378	K21631	59.7	107.7	8.20E-33	Gtl1/Pac2 family transcription factor
* XP_007604846.1hypotheticalproteinVICG_01400	K02976	24.73	84.9	5.80E-26	small subunit ribosomal protein S26e
* XP_007605182.1hypotheticalproteinVICG_01737	K02976	24.73	84.9	5.80E-26	small subunit ribosomal protein S26e
* XP_007605535.1hypotheticalproteinVICG_02090	K19613	320.8	374.1	8.60E-114	leucine-rich repeat protein SHOC2

Continued on next page

Table A2: Overview of the KEGG pathway and GO analysis of *V. corneae* HPs (Continued)

# Gene name	KO ID	Threshold	Score	E-value	KO definition
* XP_007604961.1hypotheticalproteinVICG_01515	K02934	62.5	74	1.10E-22	large subunit ribosomal protein L6e
* XP_007604991.1hypotheticalproteinVICG_01545	K03025	72.53	195.7	1.40E-59	DNA-directed RNA polymerase III subunit RPC6
* XP_007605048.1hypotheticalproteinVICG_01603	K02542	667.77	810	2.60E-245	DNA replication licensing factor MCM6 [EC:3.6.4.12]
* XP_007605092.1hypotheticalproteinVICG_01647	K14404	67.27	185	3.10E-56	cleavage and polyadenylation specificity factor subunit 4
* XP_007605096.1hypotheticalproteinVICG_01651	K02979	48.23	56.9	2.30E-17	small subunit ribosomal protein S28e
* XP_007605101.1hypotheticalproteinVICG_01656	K03681	98.07	127.7	4.60E-39	exosome complex component RRP40
* XP_007605118.1hypotheticalproteinVICG_01673	K09140	27.83	265.4	1.10E-80	pre-rRNA-processing protein TSR3
* XP_007605132.1hypotheticalproteinVICG_01687	K14326	522.83	603.1	1.10E-182	regulator of nonsense transcripts 1 [EC:3.6.4.-]
* XP_007605148.1hypotheticalproteinVICG_01703	K03165	627.8	727.7	1.90E-220	DNA topoisomerase III [EC:5.6.2.1]
* XP_007605153.1hypotheticalproteinVICG_01708	K02978	30.8	52.7	4.50E-16	small subunit ribosomal protein S27e
* XP_007605168.1hypotheticalproteinVICG_01723	K12598	1016.03	1044.1	0	ATP-dependent RNA helicase DOB1 [EC:3.6.4.13]
* XP_007605236.1hypotheticalproteinVICG_01791	K03014	89.5	114.3	7.50E-35	DNA-directed RNA polymerases I, II, and III subunit RPABC2
* XP_007605241.1hypotheticalproteinVICG_01796	K14788	197.07	206.1	9.30E-63	ribosome biogenesis protein ENP2
* XP_007605535.1hypotheticalproteinVICG_02090	K19613	320.8	374.1	8.60E-114	leucine-rich repeat protein SHOC2

Continued on next page

Table A2: Overview of the KEGG pathway and GO analysis of *V. corneae* HPs (Continued)

# Gene name	KO ID	Threshold	Score	E-value	KO definition
* XP_007605243.1hypotheticalproteinVICG_01798	K02541	598.17	732.6	7.50E-222	DNA replication licensing factor MCM3 [EC:3.6.4.12]
* XP_007605348.1hypotheticalproteinVICG_01903	K03514	176.83	227.1	4.20E-69	non-canonical poly(A) RNA polymerase PAPD5/7 [EC:2.7.7.19]
* XP_007605382.1hypotheticalproteinVICG_01937	K02883	33.3	105.8	2.40E-32	large subunit ribosomal protein L18e
* XP_007605513.1hypotheticalproteinVICG_02068	K00728	232.43	393.7	1.50E-119	dolichyl-phosphate-mannose-protein mannosyltransferase [EC:2.4.1.109]
* XP_007605535.1hypotheticalproteinVICG_02090	K19613	320.8	374.1	8.60E-114	leucine-rich repeat protein SHOC2

^a Significant scores are highlighted in the output by an asterisk.

Appendix 3: MRG Orthologs in *Microsporidia* sp. MB

Table A3: A summary of the MRG orthologs identified in *Microsporidia* sp. MB.

Gene ID	Node ID	Top BLAST Hit Description	E Value	Gene Region	Sequence Length	Protein Family
gene_1051	NODE_665_length_-5120_cov_64.6691	Cut3-like chromosome segregation protein [<i>Encephalitozoon cuniculi</i>]	0	8 .. 1111	1175	Structural maintenance of chromosomes (SMC) family member
gene_1069	NODE_681_length_-5030_cov_68.5537	DNA repair protein RAD51 [<i>Vittaforma corneae ATCC 50505</i>]	0	1 .. 331	331	RECA/RAD51/RADA DNA strand-pairing family member
gene_1101	NODE_715_length_-4904_cov_91.3836	DNA mismatch repair protein MSH2 [<i>Pancytospora philotis</i>]	0	1 .. 713	747	DNA mismatch repair protein MUTS family member
gene_114	NODE_85_length_-14989_cov_67.7995	DNA mismatch repair protein MSH6 [<i>Pancytospora epiphaga</i>]	0	3 .. 985	992	DNA mismatch repair protein MUTS family member
gene_1347	NODE_958_length_-3837_cov_70.1211	SMC3 (chondroitin sulfate proteoglycan 6) [<i>Pancytospora philotis</i>]	0	1 .. 994	1009	SMC 3
gene_1355	NODE_969_length_-3802_cov_63.0235	DNA mismatch repair protein PMS2 [<i>Pancytospora philotis</i>]	9.61E-88	1 .. 584	665	DNA mismatch repair protein MLH, PMS, MUTL
gene_941	NODE_584_length_-5545_cov_61.0005	SMC6 [<i>Enteropsectra breve</i>]	0	1 .. 951	967	SMC5, SMC6

Continued on next page

Table A3: A summary of the MRG orthologs identified in *Microsporidia* sp. MB. (Continued)

Gene ID	Node ID	Top BLAST Hit Description	E Value	Gene Region	Sequence Length	Protein Family
gene_1398	NODE_1016_length_-3626_cov_70.8003	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1 [<i>Nosema bombycis</i> CQ1]	0	6 .. 551	564	DNA annealing helicase and endonuclease ZRANB3 family member
gene_1498	NODE_1196_length_-3092_cov_68.2262	Sporulation protein RMD1 [<i>Nosema bombycis</i> CQ1]	6.07E-109	1 .. 329	330	Required for meiotic nuclear division protein 1 homolog
gene_1754	NODE_1761_length_-2146_cov_72.0349	Uncharacterized protein VICG_00967 [<i>Vittaforma corneae</i> ATCC 50505]	6.10E-28	1 .. 320	351	DNA repair and recombination proteins RAD21 /REC8
gene_1760	NODE_1784_length_-2122_cov_84.8133	DNA repair protein Rad16 [<i>Hamiltosporidium tvaerminnensis</i>]	0	4 .. 662	662	Transcription termination factor 2- related protein family
gene_1920	NODE_2705_length_-1341_cov_64.6519	MND1 [<i>Ecytonucleospora hepatopenaei</i>]	2.03E-56	1 .. 202	202	MND1
gene_421	NODE_264_length_-8941_cov_75.8837	recombination protein rad52 [<i>Vittaforma corneae</i> ATCC 50505]	7.53E-85	1 .. 183	183	DNA repair and recombination protein RAD52, RAD59
gene_478	NODE_291_length_-8441_cov_67.7299	DNA mismatch repair protein MutL [<i>Vittaforma corneae</i> ATCC 50505]	2.50E-79	18 .. 334	386	DNA mismatch repair protein MLH, PMS, MUTL
gene_603	NODE_359_length_-7512_cov_68.3672	structural maintenance of chromosome 1 [<i>Pancytospira philotis</i>]	0	1 .. 1104	1132	SMC family member
gene_684	NODE_407_length_-6941_cov_75.4349	DNA mismatch repair protein MutS [<i>Pancytospira philotis</i>]	5.58E-27	189 .. 562	379	DNA mismatch repair protein MUTS family member
gene_941	NODE_584_length_-5545_cov_61.0005	SMC6 [<i>Enteropsectra breve</i>]	0	1 .. 951	967	SMC5, SMC6

Continued on next page

Table A3: A summary of the MRG orthologs identified in *Microsporidia* sp. MB. (Continued)

Gene ID	Node ID	Top BLAST Hit Description	E Value	Gene Region	Sequence Length	Protein Family
gene_89	NODE_70_length_-16818_cov_69.0632	chromosome segregation protein [<i>Encephalitozoon cuniculi</i>]	0	1 .. 1002	1061	SMC3
gene_941	NODE_584_length_-5545_cov_61.0005	SMC6 [<i>Enteropsectra breve</i>]	0	1 .. 951	967	SMC5, SMC6

Appendix 4: Whole Genome Analysis SnakeMake Workflow

```
#This is a snakemake workflow script for conducting de novo genome
  → assembly and gene annotation
#Author: Lilian Mbaisi Ang'ang'o
#Supervisors: Prof. Ozlem Tastan Bishop and Dr. Jeremy K. Herren
#Date Created: June 11, 2023
#Research Group: Research Unit in Bioinformatics (RUBi)
#Department: Biochemistry, Microbiology and Bioinformatics
#Institute: Rhodes University
#INSTALL NECESSARY TOOLS
#QUAST
pip install quast

#BUSCO
pip install busco

#SPADES
# For Linux:
wget http://cab.spbu.ru/files/release3.15.2/SPAdes-3.15.2-Linux.tar
  → .gz
tar -xzf SPAdes-3.15.2-Linux.tar.gz

# For macOS:
brew install SPAdes

#BLAST
# BLAST can be installed as part of the NCBI BLAST+ package.
# Download the BLAST+ package from the NCBI website:
https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
```

After installation, add the BLAST executables to your PATH (
→ optional).

#KRAKEN

You need to build the Kraken database first.

Download the Kraken source code from [https://github.com/](https://github.com/DerrickWood/kraken2/releases)

→ DerrickWood/kraken2/releases

Build the database:

`kraken2-build --standard --db kraken_db`

Add the Kraken executable to your PATH (optional):

`export PATH="/path/to/kraken2:$PATH"`

#GENEMARK-ES

You need to download and install GeneMark-ES from the GeneMark

→ website:

https://exon.gatech.edu/GeneMark/license_download.cgi

After installation, add the GeneMark executable to your PATH (

→ optional):

`export PATH="/path/to/genemark/bin:$PATH"`

#BRAKER

BRAKER requires specific installations of GeneMark-ES, Augustus,

→ and other tools.

Follow the instructions in the BRAKER GitHub repository:

<https://github.com/Gaius-Augustus/BRAKER>

```

# Make sure to add all the required tools' executables to your
  ↳ PATH (optional).

#METAUK

# MetaEuk is available from GitHub.
# Download the MetaEuk repository:
git clone https://github.com/fenderglass/MetaEuk.git

# Install MetaEuk (Python and Perl dependencies are required):
cd MetaEuk
python setup.py install

# Define input and output files
reads_dir = "data"
output_dir = "results"
sample_names = ["sample"]

# Define the rule for running FastQC (Ensure read file extensions
  ↳ are similar to your dataset)
rule run_fastqc:
    input:
        reads1 = f"{reads_dir}/{sample}_R1.fastq",
        reads2 = f"{reads_dir}/{sample}_R2.fastq"
    output:
        fastqc_output = directory(f"{output_dir}/{sample}_fastqc")
    shell:
        "fastqc {input.reads1} {input.reads2} --outdir {output.
          ↳ fastqc_output}"

```

```

# Define the target rule to run the FastQC workflow for all
    ↪ samples
rule all:
    input:
        expand(f"{output_dir}/{sample}_fastqc", sample=sample_names)

# KRAKEN

import os

# Define input and output files
reads_dir = "data"
output_dir = "results"
sample_names = ["sample"]

# Define the rule for running Kraken
rule run_kraken:
    input:
        reads1 = f"{reads_dir}/{sample}_R1.fastq",
        reads2 = f"{reads_dir}/{sample}_R2.fastq",
        kraken_db = os.environ["KRAKEN_DB"]
    output:
        kraken_output = f"{output_dir}/kraken/{sample}.kraken",
        unclassified_reads = f"{output_dir}/kraken/{sample}
            ↪ _unclassified.fastq",
        taxonomy_report = f"{output_dir}/kraken/{sample}
            ↪ _taxonomy_report.txt"
    shell:
        """
        kraken2 --db {input.kraken_db} --threads 4 --paired {input.
            ↪ reads1} {input.reads2} \

```

```

        --output {output.kraken_output} --unclassified-out {output.
        ↳ unclassified_reads} --report {output.taxonomy_report}
    """

# Define the target rule to run Kraken for all samples
rule all:
    input:
        expand(f"{output_dir}/kraken/{sample}.kraken", sample=
        ↳ sample_names)

# Define the rule for genome assembly using SPAdes
rule assemble_genome:
    input:
        reads1 = f"{reads_dir}/{sample}_R1.fastq",
        reads2 = f"{reads_dir}/{sample}_R2.fastq"
    output:
        assembly = f"{output_dir}/{sample}_assembly"
    params:
        spades_params = "--careful" # You can add more parameters
        ↳ here
    shell:
        "spades.py -o {output.assembly} -1 {input.reads1} -2 {input.
        ↳ reads2} {params.spades_params}"

# Define the rule for evaluation using QUAST
rule evaluate_assembly:
    input:
        assembly = f"{output_dir}/{sample}_assembly/contigs.fasta"
    output:
        quast_report = f"{output_dir}/{sample}_quast_report"

```



```

shell:
    "quast.py -o {output.quast_report} {input.assembly}"

# Define the target rule to run the entire workflow for all
    ↪ samples
rule all:
    input:
        expand(f"{output_dir}/{sample}_quast_report", sample=
            ↪ sample_names)

# Define input and output files
assembly_file = "data/assembly.fasta"
output_dir = "results"
busco_db = "microsporidia" # Change this to the name of your BUSCO
    ↪ database
blast_db = "microsporidia_blast" # Change this to the name of your
    ↪ BLAST database

# Define the rule for running BUSCO on the assembly
rule run_busco:
    input:
        assembly = assembly_file,
    output:
        busco_output = directory(f"{output_dir}/busco")
    shell:
        "busco -i {input.assembly} -o {output.busco_output} -l {
            ↪ busco_db} -m genome"

# Define the rule for running remote BLAST on all contigs against
    ↪ the Microsporidia taxon
rule run_blast:
    input:

```

```

        assembly = assembly_file,
output:
        blast_output = directory(f"{output_dir}/blast")
shell:
        "blastn -query {input.assembly} -db {blast_db} -out {output.
        ↪ blast_output}/blast_results.txt -outfmt '6 qseqid
        ↪ sseqid evalue pident'"

# Define the target rule to run the entire workflow
rule all:
    input:
        expand(f"{output_dir}/busco/run_{busco_db}/short_summary_{
        ↪ busco_db}.txt",
            f"{output_dir}/blast/blast_results.txt")

# Define input and output files
assembly_file = "data/assembly.fasta"
output_dir = "results"

# Define the rule for running GeneMark-ES
rule run_genemark:
    input:
        assembly = assembly_file
    output:
        genemark_output = f"{output_dir}/genemark.gff"
    shell:
        "gmes_petap.pl --sequence {input.assembly} --output {output.
        ↪ genemark_output}"

# Define the rule for running Augustus
rule run_augustus:

```

```

input:
    assembly = assembly_file
output:
    augustus_output = f"{output_dir}/augustus.gff"
shell:
    "augustus_␣--species=your_species_name_␣--outfile={output.
    ↪ augustus_output}_␣{input.assembly}"

# Define the rule for running BRAKER
rule run_braker:
    input:
        assembly = assembly_file,
        genemark_output = f"{output_dir}/genemark.gff"
    output:
        braker_output = f"{output_dir}/braker.gff"
    shell:
        "braker.pl_␣--genome={input.assembly}_␣--species=
        ↪ your_species_name_␣--hints={input.genemark_output}_␣--
        ↪ cores=4"

# Define the rule for running MetaEuk
rule run_metaeuk:
    input:
        assembly = assembly_file
    output:
        metaeuk_output = f"{output_dir}/metaeuk.gff"
    shell:
        "metaeuk.py_␣-i_␣{input.assembly}_␣-o_␣{output.metaeuk_output}"

# Define the target rule to run the entire workflow
rule all:
    input:

```

```
f"{output_dir}/genemark.gff",  
f"{output_dir}/augustus.gff",  
f"{output_dir}/braker.gff",  
f"{output_dir}/metaeuk.gff"
```