THE PHYLOGENY AND EVOLUTION OF APICOMPLEXAN PARASITES

by

Varsha Mathur
B.Sc. Hons. The University of British Columbia, 2016

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

November 2020

© Varsha Mathur, 2020
The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

The phylogeny and evolution of apicomplexan parasites

submitted by Varsha Mathur in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Botany

Examining Committee:

Prof. Patrick Keeling, Department of Botany, University of British Columbia
Supervisor

Prof. Laura Wegener Parfrey, Department of Botany/Zoology, University of British Columbia
Supervisory Committee Member

Prof. Brian Leander, Department of Botany/Zoology, University of British Columbia
Supervisory Committee Member

Prof. James Berger, Department of Zoology, University of British Columbia
University Examiner

Prof. Quentin Cronk, Department of Botany, University of British Columbia
University Examiner
Abstract

Apicomplexans are a large phylum of obligate animal parasites that contain pathogens such as *Plasmodium* spp. (the causative agent of malaria) and *Toxoplasma gondii*. While these medically relevant apicomplexans are the subject of extensive research, the bulk of the diversity of the group, particularly the lineages that infect invertebrates, remain poorly studied and largely ignored in high-throughput sequencing surveys. In this dissertation, I show that these groups are critical to gaining insights into the origins and evolution of the Apicomplexa. I begin by examining the diversity and inferred ecology of the enigmatic apicomplexan-related lineages (ARLs), and show that ARL-V is highly abundant in environmental surveys, and is tightly associated with coral tissue and mucus, suggesting that it represents a core symbiont of coral. In the following chapters, using methods of single-cell transcriptomics, I sequenced the transcriptomes of 15 invertebrate-infecting apicomplexans. Using this dataset, I constructed a robust and taxon-rich multi-gene apicomplexan phylogeny that resolves the deep phylogenetic relationships within the group, and also form a new class of apicomplexans, the Marosporida, that is sister to the Hematozoa and Coccidia. Most unexpectedly, in Chapter 2, I show that certain taxa previously classified as apicomplexans, actually represent convergently evolved animal parasites, suggesting that apicomplexan-like parasites have evolved at least four times independently. In Chapter 3, I examine the presence and function of apicoplasts (remnant plastids) across the diversity of the group using whole genome shotgun sequencing (WGS), and find that the Marosporida contain the smallest, most AT-rich, and gene poor apicoplast genomes sequenced to date. I also present the first evidence of plastids in the gregarines, and show that archigregarines retain the canonical apicomplexan plastid metabolism, whereas only one clade of marine eugregarines retains plastids that solely carry out type II fatty acid biosynthesis. Lastly in Chapter 4, I reconstruct the mitochondrial metabolism in the gregarines and squirmids, and find that eugregarines contain highly reduced respiratory chains, suggesting that they have lost their mitochondrial genomes, and possess limited energy metabolism. Altogether, the data presented here, illustrates the significance of invertebrate-infecting apicomplexans in illuminating the early evolution of the apicomplexans and myzozoans.
Lay Summary

Apicomplexans are obligate parasites of animals that cause devastating diseases such as malaria, toxoplasmosis, and cryptosporidiosis. Our understanding of how these parasites evolved remains incomplete because we currently lack genetic data from most of the diversity of the group, particularly from the apicomplexans that infect invertebrates. In this dissertation, I aimed to fill these knowledge gaps, and sequenced the expressed genes of diverse invertebrate-infecting apicomplexans. Using this comprehensive dataset, I reconstructed the evolutionary relationships between apicomplexan lineages, and examined the evolution and distribution of fundamental cellular features that define the group. Overall, this research provides new insights into the evolutionary history of the largest group of eukaryotic parasites on earth.
Preface

The research conducted in this thesis has been published or submitted to peer-reviewed journals and was carried out in collaboration with other scientists. Below I detail the contributions and publications corresponding to each thesis chapter.

Chapter 2: Global diversity and distribution of close relatives of apicomplexan parasites
This study was originally conceived and designed as an extension of a pilot project carried out in my bachelor’s degree honours thesis. The project was designed by Martin Kolisko, Javier del Campo, Patrick Keeling, and I. I conducted all bioinformatics analyses with guidance from Martin Kolisko and Javier del Campo. I wrote the manuscript and created the figures and datasets with input from all of the authors. A version of this chapter has been published as:


Chapter 3: Multiple independent origins of apicomplexan-like parasites
This project was conceived and designed by Patrick Keeling, Martin Kolisko and I. Árni Kristmundsson, Mark Freeman, Brian Leander, and I collected the samples. Martin Kolisko and I carried out the transcriptome and genome sequencing preparation. Martin Kolisko, Nicholas Irwin, Elisabeth Hehenberger, and I carried out the phylogenomics and plastid related analyses. Martin Kolisko and I carried out all other analyses. I designed all the figures. Patrick Keeling and I jointly wrote the manuscript with input from all authors. A version of this chapter has been published as:


* Authors contributed equally to the publication.
Chapter 4: Plastid evolution in deep-branching apicomplexans

Chapter 4 was conceived and designed by Patrick Keeling and I. Árni Kristmundsson, Mark Freeman and Camino Gestal collected the samples and isolated the parasites. Filip Husnik and I prepared the genome and transcriptome sequencing libraries. Waldan Kwong and I assembled and annotated the plastid genomes. I carried out all other data analyses and designed the figures. I wrote the manuscript with input from all other authors. The manuscript has been accepted as:


Chapter 5: Organellar metabolism in the gregarines and squirmids

Chapter 5 was conceived and designed by Patrick Keeling and I. Kevin Wakeman and I collected the samples and isolated the parasites. I prepared the samples for transcriptome sequencing. I carried out all other bioinformatic analyses with assistance from Martin Kolisko, Racquel Singh and Morelia Trznadel.
# Table of Contents

Abstract ........................................................................................................................................ iv
Lay Summary ............................................................................................................................... vii
Preface .......................................................................................................................................... v
Preface .......................................................................................................................................... v
Table of Contents ........................................................................................................................ vii
List of Tables ............................................................................................................................... xii
List of Figures ............................................................................................................................. xiii
List of Abbreviations ................................................................................................................ xiv
Acknowledgements ................................................................................................................... xvi
Dedication ..................................................................................................................................... xvii

## Chapter 1: Introduction ............................................................................................................. 1

1.1 The significance of apicomplexans ....................................................................................... 1

1.2 Defining cellular and molecular features of apicomplexans .............................................. 2

   The apical complex .................................................................................................................. 2

   The glideosome ...................................................................................................................... 4

   The apicoplast ....................................................................................................................... 4

      1.2.1.1 The evolutionary history of the apicoplast ............................................................ 4

      1.2.1.2 The apicoplast genome and function ......................................................................... 5

      1.2.1.3 Loss of the apicoplast in Cryptosporidium spp. parasites ......................................... 6

   The apicomplexan mitochondria ............................................................................................ 7

1.3 Diversity and classification of apicomplexans .................................................................... 7

   Relatives of apicomplexans: the SAR supergroup, alveolates and myzozoans .................. 7

   Current state of the taxonomy of apicomplexans ................................................................. 8

   The gregarines ....................................................................................................................... 9

1.4 The origin of apicomplexans parasites ................................................................................. 10

   The chromerids and colpodellids ......................................................................................... 10

   The discovery of apicomplexan-like lineages (ARLs) ......................................................... 11

1.5 Research Objectives ............................................................................................................ 13
Chapter 2: Global diversity and distribution of close relatives of the apicomplexan parasites .......................................................... 15

2.1 Introduction ............................................................................. 15

2.2 Results and Discussion ......................................................... 17

Distribution of ARLs confirms association with coral reefs .......... 17
Chromera and Vitrella are not coral symbionts ....................... 19
Refining ARL-V and its association with coral ....................... 21

2.3 Chapter 2 conclusions ............................................................ 24

2.4 Materials and methods .......................................................... 25

Amplicon sequence retrieval .................................................. 25
Selecting and clustering plastid 16S rRNA gene sequences .......... 25
Building reference phylogeny and alignment ......................... 26
Short reads assignment and EPA tree reconstruction using QIIME .... 26
QIIME analysis: operational taxonomic unit (OTU) picking ........ 26
Ecological meta-analyses ........................................................... 27

2.5 Chapter-specific acknowledgments ........................................ 27

Chapter 3: Multiple independent origins of apicomplexan-like parasites ......................................................... 28

3.1 Introduction ............................................................................. 28

3.2 Results and Discussion ......................................................... 28

3.3 Chapter 3 conclusions ............................................................ 35

3.4 Materials and Methods .......................................................... 35

Amplicon Genomics and transcriptomics of Piridium sociabile .... 35
Genome assembly and annotation of Piridium sociabile ........... 36
Transcriptomics of the gregarines and Platyproteum .............. 36
Transcriptome assembly and annotation of the gregarines and Platyproteum ............................................................ 37
Ortholog identification, gene concatenation and phylogenomics .... 37
Search and identification of plastid proteins ......................... 38
Search for plastid localization signals .................................. 40
Analysis of cellular metabolic pathways ................................ 40

3.5 Chapter-specific acknowledgments ........................................ 41
Chapter 4: Plastid evolution in deep-branching apicomplexans ................................................. 42

4.1 Introduction .................................................................................................................. 42

4.2 Results and Discussion: ............................................................................................ 44

A resolved multiprotein phylogeny of the Apicomplexa .................................................. 44
A new apicomplexan class, Marosporida, that infects marine invertebrates ................. 45
Taxon sampling does not improve congruence between apicoplast and nuclear phylogenies .......................................................................................................................... 48

Apicoplast genomes in Margolisiella and Rhytidocystis are highly reduced and lack RNA polymerase genes ................................................................. 50
The enigmatic transcription of Margolisiella and Rhytidocystis plastid genes ............ 53
The canonical apicoplast biosynthetic function is conserved in Marosporida ............ 57

4.3 Chapter 4 conclusions ................................................................................................. 58
Taxonomic description .................................................................................................. 59

4.4 Materials and Methods .............................................................................................. 59
Sample collection and DNA/RNA extraction ............................................................... 59
Genome sequencing, plastid genome assembly, and annotation .................................. 60
Plastid phylogenomic analyses ..................................................................................... 61
Transcriptome sequencing and assembly ....................................................................... 62
Nuclear phylogenomics analyses .................................................................................. 63
RNA polymerase analysis ............................................................................................. 64
Search for plastid-derived biosynthetic proteins ........................................................... 64
Mitochondrial and 18S ribosomal small subunit gene phylogenies ................................. 65

4.5 Chapter-specific acknowledgments ............................................................................. 66

Chapter 5: Organellar metabolism in the gregarines and squirmids ............................... 67

5.1 Introduction .................................................................................................................. 67

5.2 Results and Discussion: ............................................................................................ 69
Phylogenetic placement of Squirmida confirmed with additional taxon sampling .... 69
Phylogenomics refines taxonomic relationships within the gregarines ...................... 71
Gregarines are monophyletic with the exclusion of Cryptosporidium ......................... 72
Squirmid plastid metabolism is highly convergent to apicomplexans ......................... 72
Multiple losses of plastids in the gregarines

The eugregarines possess a patchy distribution of ETC components

Reconstruction of the flagellar apparatus in the gregarines and squirmids

5.3 Chapter 5 conclusions

5.4 Materials and Methods

Conclusions

Chapter summaries and conclusions

Future directions

References

Appendices

Appendix A - Supplementary Information for Chapter 2

Appendix B - Supplementary Information for Chapter 3

Appendix C - Supplementary Information for Chapter 4

Figure A3. Outlier removal for distribution analysis

Table A1. List of Illumina and 454 amplicon sequencing surveys analysed

Figure B1. Myzozoan 18S SSU rRNA phylogenetic tree

Figure B2. Plastid phylogenomic tree

Table B1. Plastid-targeted proteins and localization signal analysis

Table B2. Presence of Metabolic Pathways based on KEGG Modules

Figure C1. Maximum likelihood (ML) 18S rRNA gene phylogeny of the Apicomplexa with dinoflagellates as the outgroup

Figure C2. Phylogenetic placement of Aggregata, Merocystis, and Margolisiella based on mitochondria-encoded proteins
Figure C3. Phylogenetic placement of *Aggregata*, *Merocystis*, and *Margolisiella* based on 22 plastid-encoded proteins .......................................................... 124

Figure C5. Complete circular apicoplast genomes ................................ 126

Table C1. Nuclear-encoded phylogenomic statistics........................ 127

Table C2. Plastid phylogenomics statistics .................................... 128

Table C3. Predicted plastid-targeted proteins .................................. 129
List of Tables

Table 1. Reconstruction of the gregarine and squirmid mitochondrial metabolism based on KEGG..............................................................78
Table A1. List of Illumina and 454 amplicon sequencing surveys analysed........................................116
Table B1. Plastid-targeted proteins and localization signal analysis........................................119
Table B2. Presence of Metabolic Pathways based on KEGG Modules........................................118
Table C1. Nuclear-encoded phylogenomic statistics............................................................126
Table C2. Plastid phylogenomics statistics........................................................................127
Table C3. Predicted plastid-targeted proteins.....................................................................128
List of Figures

Figure 1: The diversity of animal symbiosis in the Apicomplexa.................................2
Figure 2: Diagram of apicomplexan trophozoite based on studies of Toxoplasma gondii........3
Figure 3: Current state of molecular data available for apicomplexans and chrompodellids......12
Figure 4: Maximum likelihood (ML) 18S rRNA gene phylogeny of the gregarines...............14
Figure 5: Maximum likelihood phylogenetic tree inferred from the plastid 16S rRNA gene......18
Figure 6: The proportion of Apicomplexan-Related Lineages (ARLs) recovered from specific coral reef environments...............................................................20
Figure 7: The absolute abundances of the Apicomplexan-Related Lineages (ARL) sequences recovered in 24 coral species.................................................................22
Figure 8: Phylogenomic tree of the Apicomplexa and relatives........................................30
Figure 9: Plastid dependency in Piridium and Platyproteum has evolved convergently to apicomplexans.........................................................................................32
Figure 10: The distribution of cellular metabolic pathways across the tree of apicomplexans and chrompodellids..............................................................................34
Figure 11: Light micrographs of oocysts of the species sequenced......................................44
Figure 12: Phylogeny of the Apicomplexa and placement of the new class, Marosporida......47
Figure 13: Apicoplast-encoded gene phylogeny.................................................................49
Figure 14: The apicoplast genomes of Aggregata octopiana, Merocystis kathae, and Margolisiella islandica.................................................................51
Figure 15: The RNA polymerases (RNAPs) present in plastid-bearing apicomplexans, Digyalum, and Piridium.................................................................55
Figure 16: Multiprotein phylogeny of the Apicomplexa, chrompodellids and Squirmida....71
Figure 17: Metabolic reconstruction of isoprenoid, fatty acid, heme, and iron-sulphur cluster biosynthesis in the gregarines and squirmids.................................74
Table 1: Reconstruction of the gregarine and squirmid mitochondrial metabolism based on KEGG........................................................................................................78
Figure 18: Flagellar apparatus reconstruction across apicomplexans, chrompodellids and squirmids.........................................................................................81
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARLs</td>
<td>Apicomplexan-related lineages</td>
</tr>
<tr>
<td>AT</td>
<td>Adenine thymine</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic local alignment search tool</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>BUSCO</td>
<td>Benchmarking universal single copy orthologues</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary deoxyribonucleic acid</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding sequence</td>
</tr>
<tr>
<td>C-terminal</td>
<td>Carboxy-terminus</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>G4</td>
<td>Gamma distributed rates with four categories</td>
</tr>
<tr>
<td>GC</td>
<td>Guanine cytosine</td>
</tr>
<tr>
<td>GTR</td>
<td>Generalised time-reversible model</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov model</td>
</tr>
<tr>
<td>KAAS</td>
<td>KEGG Automatic Annotation Server</td>
</tr>
<tr>
<td>kbp</td>
<td>Kilo base pair</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>MAFFT</td>
<td>Multiple alignment using fast Fourier transform</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov chain Monte Carlo</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum likelihood</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>N-terminal</td>
<td>Amino-terminus</td>
</tr>
<tr>
<td>OTUs</td>
<td>Operational taxonomic unit with 97% clustering threshold</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PP</td>
<td>Posterior probability</td>
</tr>
<tr>
<td>QIIME</td>
<td>Quantitative Insights into Microbial Ecology</td>
</tr>
<tr>
<td>R8</td>
<td>Free rate model with eight categories</td>
</tr>
<tr>
<td>RAxML</td>
<td>Randomized accelerated maximum likelihood</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>RNAP</td>
<td>RNA polymerase</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal ribonucleic acid</td>
</tr>
<tr>
<td>SAR</td>
<td>Stramenopiles, Alveolates, Rhizarians</td>
</tr>
<tr>
<td>SCaFoS</td>
<td>Selection, concatenation, and fusion of sequences</td>
</tr>
<tr>
<td>SSU</td>
<td>Small subunit ribosomal gene</td>
</tr>
<tr>
<td>TIC</td>
<td>Translocase of the inner chloroplast membrane</td>
</tr>
<tr>
<td>TOC</td>
<td>Translocase of the outer chloroplast membrane</td>
</tr>
<tr>
<td>TP</td>
<td>Transit peptide</td>
</tr>
<tr>
<td>tRNA</td>
<td>Transfer ribonucleic acid</td>
</tr>
</tbody>
</table>
Acknowledgements

I would like to express my sincere gratitude to my PhD and BSc. thesis supervisor, Patrick Keeling. I started working in the Keeling lab in September 2014 as a clueless undergraduate student. Thank you for giving me a chance back then, and thank you for all the opportunities you have given me since. It is in the Keeling lab that I became a scientist, and for that I will always be so grateful.

Thank you to my mentors and co-authors over the years: Martin Kolisko, Javier del Campo, Filip Husnik, and Waldan Kwong. I am so thankful for their patience, encouragement, and all that I learnt from these brilliant scientists. Thank you to Kevin Wakeman for collecting and dissecting the most gruesome beasts with me, without your spirit I could not have achieved the goals of this thesis. A huge thank you to all the Keeling lab members, past and present, who provided the most enjoyable and positive lab atmosphere, it was truly a pleasure coming into the lab every day. I would also like to thank my supervisory committee members, Laura Parfrey and Brian Leander, for their guidance and support.

Finally, to my wonderful family, Rahul, Priya and Meghna, thank you for empowering me to follow my dreams, cheering me on from the other side of the world, and always being there for me. Thank you to the Irwin family, Dave, Kitty, Liam and Daniel, for being a huge support and my family away from home. Thank you to my best friends, Meha, Racquel, and Veronica, who have been by my side through all the highs and lows of these four years. And last but not least, thank you to my partner, Nick, for inspiring me every single day.
To my mother, Priya.

"I am an example of what is possible when girls from the very beginning of their lives are loved and nurtured by people around them."

Michelle Obama
Chapter 1: Introduction

The research presented here focuses on the origins and evolution of apicomplexan parasites. Apicomplexans are a diverse group of obligate animal parasites that are responsible for causing many medically and economically relevant diseases. While the pathogenic parasites are the focus of most apicomplexan research, the majority of the diversity of the phylum, particularly those infecting invertebrates, are extremely understudied and represent a promising opportunity to gain insights into the evolutionary history of this group. In this introductory chapter, I summarize the significance and defining features of apicomplexans and their relatives. I outline relevant advances in the field and identify key research questions focused towards gaining a resolved apicomplexan phylogeny and a better understanding of apicomplexan evolution.

1.1 The significance of apicomplexans

The Apicomplexa are a large eukaryotic phylum of obligatory parasites. They specifically infect animals and there are more than 6000 named species that infect both invertebrates and vertebrates (Votýpka et al. 2017). Apicomplexans are the causative agents of serious diseases, the most devastating being the malaria parasite (Plasmodium spp.). In 2018, there were an estimated 228 million cases of malaria worldwide that disproportionately impacted poor communities and children in developing countries (www.who.int/malaria). Apicomplexans also impact human livelihood by causing diseases in livestock, such as the tickborne East Coast cattle fever (Theileria parva) and poultry coccidiosis (Eimeria spp.). These parasites have shaped the course of human history and evidence of malaria plagues are even found in 5,000 year old Egyptian papyri, and historical records of the Greeks and Romans (Cox 2010).

While most well-recognized apicomplexan taxa cause diseases, a large diversity of this phylum are non-pathogenic, and recent studies show that they can even be mutualists, for example, in the case of the tunicate symbiont, Nephromyces (Saffo et al. 2010). Therefore, apicomplexans often affect their hosts in diverse ways along the symbiotic spectrum, and a deeper investigation of apicomplexan diversity in recent years has revealed more complex interspecific relationships than exclusively parasitism within the group (Rueckert et al. 2019) (Figure 1).
1.2 Defining cellular and molecular features of apicomplexans

The apical complex

Irrespective of the nature of the symbiosis with their hosts, apicomplexans have a highly specialized biology that allows them to invade and survive within an animal host. Key to their success in doing so, and the namesake of the Apicomplexa, is the apical complex, a specialized structure at the anterior end of the cell. Functional studies of the apical complex stem from select experimentally amenable taxa, namely Toxoplasma and Plasmodium, therefore our mechanistic understanding of this unique cell structure and its conservation across the phylum is still limited (Simdyanov et al. 2017). The apical complex consists of three distinct groups of components: the apical cap which is formed by the inner membrane complex (IMC) (or alveoli, see section 1.3.1) that extends along the length of the parasite, the conoid which comprises a cone of spiraling tubulin-rich fibers, and secretory organelles known as the micronemes and rhoptries (Frénal et al. 2017; Dos Santos Pacheco et al. 2020). When an apicomplexan invades a host cell, the apical complex is the site of exocytosis of initial host-recognition and -binding molecules that allow for penetration of the host plasma membrane (Kats et al. 2007; Lebrun et al. 2013).
Although the apical complex is the unifying morphological feature of the phylum, different groups utilize it in distinct ways. In the case of intracellular parasites, such as *Toxoplasma* and *Plasmodium*, upon host cell invasion, molecules released from the secretory organelles are used to modify the host cell environment and form a parasitophorous vacuole (which is derived from the invagination of the host cell plasma membrane) to protect the parasite from the host cell defences (Kemp et al. 2013). However, in other taxa such as *Cryptosporidium* the host is only partially penetrated and the apicomplexan attaches onto a host cell via an embedded apical complex, or in some gregarines, the apicomplexan remains extracellular with the host cell remaining virtually unaltered (Schrével et al. 2016; Simdyanov et al. 2017).

Interestingly, the apical complex has evolved from an ancient cell structure used for a mode of feeding known as myzocytosis; a process in which the predator punctures the prey cell membrane and sucks out the cellular content (Schnepf and Deichgräber 1984). This unique feeding strategy of 'cellular vampirism' is seen today in close relatives of the apicomplexans, such as colpodellids and some dinoflagellates (Leander, Kuvardina, et al. 2003; Okamoto and Keeling 2014). The ancestral capability of myzocytosis in these related groups has led to their collective name, the Myzozoa (Cavalier-Smith and Chao 2004) (See section 1.3.1).

**Figure 2:** Diagram of apicomplexan trophozoite based on studies of *Toxoplasma gondii*. Key apicomplexan subcellar compartments and structures. Adapted from (Barylyuk et al. 2020).
The glideosome

The glideosome is another unique apicomplexan machinery used for a type of movement known as gliding motility, that is distinctive from the amoeboid motion involving pseudopods, and from the ciliary/flagellar motion used by most protists (Opitz and Soldati 2002). Gliding motility is used by the motile and/or invasive stages of the apicomplexan life cycle (sporozoites, merozoites and oocinetes), generally called 'zoites', to migrate across biological barriers (such as hosts cells and tissues), as well as for active host cell entry and egress from infected cells (Heintzelman 2015). Similar to the apical complex, the glideosome has been studied in the most experimentally tractable taxa, *Toxoplasma* and *Plasmodium*, and is composed of an actin-myosin based motor apparatus that is located underneath the plasma membrane and promotes substrate-dependent gliding (Frénal et al. 2017). A key role in gliding motility is played by the IMC (the alveoli) of the cell as it provides an anchor for the movement. The micronemes also play an important role in releasing adhesins that attach to the cytoskeleton and mediate interaction with extracellular host cell receptors to propel the apicomplexan zoite forwards (Paing and Tolia 2014). Although the specific components of the glideosome were discovered over the past decade, and continue to be re-examined, the 'capping model' of gliding motility in Apicomplexa was conceptually formulated more than a century ago, and is now substantiated as a host–parasite interacting complex that is translocated from the front to the rear of the parasite by the action of the glideosome (Frénal et al. 2017)

The apicoplast

1.2.1.1 The evolutionary history of the apicoplast

Another defining feature of the Apicomplexa is their non-photosynthetic plastid (chloroplast), known as the apicoplast. The discovery of a remnant plastid in a group of obligate parasites was extremely surprising (Wilson et al. 1996) and has since been heavily researched as a potential drug target for apicomplexan diseases (McFadden and Roos 1999; Ralph et al. 2001). The presence of this organelle also revealed an important aspect about the evolutionary history of the group, and shows that apicomplexans evolved from free-living, photosynthetic ancestors.
The apicoplast has four surrounding membranes indicating that it is of secondary endosymbiotic origin (Lim and McFadden 2010), which refers to a process in which a phagotrophic eukaryote engulfs and retains another eukaryote with a plastid obtained by primary endosymbiosis of a cyanobacterium-like prokaryote (Keeling 2010). Secondary endosymbiotic plastids typically have three or four membranes, whereas primary plastids have only two membranes, which are homologous to the two membranes of the ancestral cyanobacteria-like endosymbiont (Keeling 2010). The outermost membrane of the apicoplast is analogous to the phagosomal membrane of the host cell, and the second outermost membrane (periplastid membrane) of the apicoplast originates from the plasma membrane of the engulfed alga cell (Lim and McFadden 2010). While the evolutionary origin of the apicoplast was previously contentious (Keeling 2010), it has now been definitively shown to a red-algal derived plastid (Janouškovec et al. 2010) (See section 1.4.1).

1.2.1.2 The apicoplast genome and function

Compared to typical photosynthetic chloroplast genomes, the compact ~35 kbp apicoplast genomes are some of the smallest known to date (Su et al. 2019). Across apicomplexan lineages, the gene content of apicoplasts has proven to be remarkably conserved; they have lost all genes encoding proteins that function directly in photosynthetic electron transfer (i.e., photosynthesis-related genes) and their genomes are thought to be retained due to the retention of a small number of non-housekeeping genes (Janouškovec et al. 2010).

Despite the loss of photosynthesis, the apicoplast is indispensable and carries out four essential metabolic functions: the biosynthesis of fatty acids (via the FASII pathway), isoprenoids (via the MEP/DOXP pathway), iron-sulphur clusters, and heme (or tetrapyrroles) (Lim and McFadden 2010). The genes functioning in these pathways have been transferred to the nuclear genome which is a hallmark feature of endosymbiosis. The nucleus-encoded apicoplast proteins are then imported back into the organelle using a bipartite sequence leader consisting of a canonical eukaryotic signal peptide (SP) followed by a plastid-derived transit peptide (TP) at the N-terminus of the protein (Waller et al. 1998; Waller 2000). The SP mediates import into the endomembrane system and the TP mediates sorting and trafficking to the apicoplast (Boucher
and Yeh 2019). Once the protein is delivered to the outermost apicoplast membrane, it crosses the periplastid membrane using a repurposed ER-associated degradation pathway (ERAD), which is a conserved eukaryotic pathway typically used for retro-translocating misfolded proteins from the ER to the cytoplasm for degradation by the ubiquitin–proteasome system (Boucher and Yeh 2019). The final protein translocation across the inner two apicoplast membranes takes place using the canonical plastid TOC and TIC protein complexes of primary plastids (Van Dooren et al. 2008; Glaser et al. 2012; Sheiner et al. 2015).

1.2.1.3 Loss of the apicoplast in Cryptosporidium spp. parasites

The metabolic pathways that drive the retention of plastids, even with the loss of photosynthesis, are typically present in the eukaryote before the gain of the plastid. Therefore, endosymbiosis leads to a metabolic redundancy that often causes the loss of many original host-derived pathways, making the plastidial version of the pathway indispensable. Hence, loss of a plastid is an extremely difficult evolutionary step that can only occur if the organism can avoid or overcome this dependency, either by maintaining the cytosolic version of the pathway, scavenging the metabolite from its environment, or eliminating the requirement of the metabolite altogether (Gornik et al. 2015). Only two eukaryotic lineages have been able to disentangle from this complex plastid dependency, one of which is the apicomplexan, Cryptosporidium, which has completely lost the apicoplast (Abrahamsen et al. 2004; Xu et al. 2004; Bessoff et al. 2013). Cryptosporidium has achieved this by drastically reducing its requirement for heme (has only one known cytochrome), and can scavenge isoprenoids and most other metabolites from its host (Van Dooren et al. 2012; Zhu and Guo 2014). It has also been proposed that the single host life cycle and extracellular infection of Cryptosporidium parasites might have contributed to overcoming plastid dependency (Gornik et al. 2015). There are conflicting opinions on whether gregarines, that are proposed to be the closest relatives of Cryptosporidium, possess a plastid or not. However, they remain relatively understudied compared to the medically significant apicomplexans, and based on the scarce genome sequencing data available from the group, are thought to have lost their plastid as well (Toso and Omoto 2007; Templeton et al. 2010) (See section 1.3.3).
The apicomplexan mitochondria

Apicomplexans also have an unusual mitochondria with tubular cristae that contain a tiny genome ranging from 5.8 kbp in *Parahaemoproteus* to 11.0 kbp in *Babesia*, and can either be composed of monomeric linear chromosomes or circularly permuted linear arrays (Feagin et al. 1997). Three protein-coding genes (*cox1, cox3, and cob*) and fragments of the small and large ribosomal rRNA genes are conserved in the mitochondrial genomes across the group. With the exception of the DNA-lacking mitochondria of *Cryptosporidium*, which has a reduced, relic organelle resembling the mitosomes of microsporidia and diplomonads (Keeling 2004). Across eukaryotes, four respiratory complexes (I-IV) and the proton-driven ATP-synthase are fundamental to the oxidative phosphorylation pathway central to mitochondrial metabolism, but these pathways can be lost in organisms adapted to anaerobic or sugar-rich environments, often as a consequence of parasitism (Muller et al. 2012). The apicomplexans are unique, however, in that even aerobic species have lost respiratory complex I (Sheiner et al. 2013).

1.3 Diversity and classification of apicomplexans

*Relatives of apicomplexans: the SAR supergroup, alveolates and myzozoans*

For decades, due to the absence of a fossil record and highly divergent cellular characteristics, it was difficult to place the Apicomplexa within the larger context of eukaryotic diversity. However, with the advent of molecular studies in the 1990s, sequencing of the 18S small subunit (SSU) rRNA gene provided the first phylogenetic evidence of their relationship with other eukaryotes, specifically to the ciliates and dinoflagellates (Gajadhar et al. 1991; Wolters 1991). This reinforced the suspected homology of the flattened subpellicular sacs observed in ultrastructural studies under the plasma membrane of all three lineages, known as the cortical alveoli. Among the alveolates, apicomplexans are more closely related to the dinoflagellates and are united by the ancestral ability to feed via myzocytosis (see section 1.2.1) (Fast et al. 2002). The monophyly of the Alveolates (apicomplexans, dinoflagellates and ciliates) is now supported with multiprotein phylogenies, and are also shown to be more broadly related to the Stramenopiles and Rhizarians, together forming the clade 'SAR' (Burki et al. 2020). The SAR 'supergroup' has been estimated to comprise almost half of all extant eukaryotic diversity including microbial algae (e.g. diatoms and dinoflagellates), ecologically important free-living
protists (e.g. foraminiferans and radiolarians), and devastating parasites (e.g. apicomplexans and oomycetes) (Del Campo et al. 2014).

**Current state of the taxonomy of apicomplexans**

Historically, classification of taxa within the Apicomplexa was made based on microscopic and histological observations of life cycle stages, cell morphology, and host(s) identity. The phylum was split into two subphyla based on whether the asexual stage possesses a conoid or not; the Aconoidasida (or Hematozoa) (Mehlhorn et al. 1980; Vivier 1982) and Conoidasida (Levine, 1988). The monophyly of the Hematozoa has now been substantiated by both 18S SSU gene phylogenies as well as multi-protein phylogenomics, and consists of the classes Haemosporidia (includes the malarial parasites *Plasmodium* spp.), Piroplasmida (e.g. *Babesia*, *Theileria*) and Nephromycida (the tunicate mutualist, *Nephromyces*) (Muñoz-Gómez et al. 2019).

The subphyla Conoidasida contains the Coccidia (e.g. *Toxoplasma*, *Eimeria*) and the Gregarinasina, commonly referred to as the gregarines (Adl et al. 2019). However, this group is paraphyletic and molecular phylogenies show that the Coccidia are sisters to the Hematozoa to the exclusion of gregarines (Janouškovec et al. 2015). The Haemosporidia, Piroplasmida, and Coccidia contain the bulk of the medically and economically important parasites, and therefore have been the focus of most apicomplexan research. In the past decade, with the explosion of high-throughput sequencing technologies, there have been significant genome sequencing surveys, transcriptomics, proteomics, and other cutting-edge experimental work done on these groups, and in particular on the experimentally amenable taxa, *Toxoplasma* and *Plasmodium* (Solyakov et al. 2011; Sidik et al. 2016; Koreny et al. 2020; Dos Santos Pacheco et al. 2020). Consequently, the evolutionary relationships between and within these groups are well-resolved and substantiated by mulitgene and well-sampled phylogenies (Kuo et al. 2008; Burki et al. 2016; Swapna and Parkinson 2017). However, the gregarines, and other deep-branching apicomplexan lineages (unstudied groups classified as 'incertae sedis') are greatly lacking in high-throughput sequencing data (Leander 2008; Morrison 2009). This limits our understanding of apicomplexan evolution and the origin of parasitism itself. And in turn, hinders insights into the structural and molecular mechanisms that underpin apicomplexan infection such as the apical
complex, glideosome, and how these relate to the parasite life cycles, host specificity, and habitats.

**The gregarines**

Gregarines represent an extremely large and abundant group of early-branching apicomplexans that exclusively infect invertebrate hosts. They are highly ubiquitous in both terrestrial and marine environments, and found in diverse hosts including annelids, molluscs, echinoderms, sipunculids and arthropods (Desportes and Schrével 2013). Gregarines have monoxenous (single host) life cycles and generally attach to the epithelial tissue in the gut lumen of the host. However, some species can be found in coelomic cavities and tissues associated with the reproductive system (Leander 2008; Desportes and Schrével 2013). Recent metagenomic (metabarcoding) studies exploring the eukaryotic diversity in terrestrial ecosystems have shown a high diversity and dominance of gregarines in soils, revealing the key role play in terrestrial environments (Mahé et al. 2017; Lentendu et al. 2018).

Differing views concerning the taxonomy of the gregarines are emerging based on 18S rRNA gene phylogenies (Cavalier-Smith 2014; Simdyanov et al. 2017). The 18S rRNA gene in gregarines is extremely fast-evolving and problematic for phylogenetic analysis, particularly in resolving deeper nodes due to artifacts related to long branch attraction (LBA) (Figure 4) (Simdyanov et al. 2015). Therefore, an overall comprehensive and reliable taxonomic review is still missing for this group. The most recent review of eukaryotes still refers to the three traditional gregarine groups, Archigregarinorida, Eugregarinorida, and Neogregarinorida, that are classified based on habitat, host range, and trophozoite (feeding stage) morphology, and is, at least to some extent, supported by 18S ribosomal gene phylogenies (Iritani et al. 2018; Adl et al. 2019). The archigregarines are the most ancestral group and infect only marine invertebrates (Rueckert and Leander 2009; Rueckert and Horák 2017), eugregarines can be found in marine, freshwater, and terrestrial habitats, and infect the intestines, coeloms, and reproductive vesicles (Leander 2008), and neogregarines are found in terrestrial environments, primarily infecting insects, and develop intracellularly in the host tissue (Votýpka et al. 2017). The vertebrate-infecting *Cryptosporidium* was traditionally classified in the Coccidia, however, phylogenies
based on the 18S rRNA and beta-tubulin genes suggested that Cryptosporidium forms a separate and distinct clade from the Coccidia, and are putatively sister to the gregarines (Carreno et al. 1999; Leander, Clopton, et al. 2003). However, a paucity of genome or transcriptome data from gregarine taxa impedes conclusive evidence substantiating the monophyly of Cryptosporidium and the gregarines.

1.4 The origin of apicomplexans parasites

The chromerids and colpodellids

Questions regarding apicomplexan origins have been of interest in general and after the discovery of the apicoplast, the question became even more precise and fascinating, how did a photosynthetic ancestor turn into an obligate symbiont of animals? Critical to approaching this question was the discovery of close photosynthetic relatives of the apicomplexans, Chromera velia and Vitrella brassicaformis in coral reefs (Moore et al. 2008; Janouškovec et al. 2010; Oborník et al. 2012). Chromera and Vitrella, known as the chromerids, are related to the non-photosynthetic and phagotrophic colpodellids in a paraphyletic assemblage (Figure 3). This group of free-living autotrophs and phagotrophs, commonly referred to as the 'chrompodellids', are the sister group to the Apicomplexa (Kuvardina et al. 2002; Gile and Slamovits 2014; Janouškovec et al. 2015).

Genome and transcriptome sequencing of Chromera and Vitrella allowed for key insights into the apicomplexan ancestor and the cellular and genomic changes that underlie the transition to parasitism (Flegontov et al. 2015; Janouškovec et al. 2015; Woo et al. 2015). By reconstructing the gene repertoire of the apicomplexan ancestor, these studies found that proteins from key metabolic pathways and the endomembrane trafficking systems have been lost during the transition to obligate symbiosis (Janouškovec et al. 2015; Woo et al. 2015). They show that the apicomplexan ancestor contained a broad repertoire of genes and many of which were modified for infection processes, such as extracellular proteins, components of a motility apparatus, and DNA- and RNA-binding protein families (Woo et al. 2015). Therefore, they proposed that the evolution of obligate animal symbiosis may not have been driven by novel 'parasite innovations', but rather loss and repurposing of the existing gene repertoire (Janouškovec et al. 2015).
Furthermore, the sequencing of the chloroplast genomes of *Chromera* and *Vitrella* was also instrumental in revealing the origin of the myzozoan plastid. Previous molecular data led to conflicting conclusions supporting either its green algal origin or red algal origin, which was exacerbated by non-overlapping plastid genes in dinoflagellates and apicomplexans making the two plastids nearly impossible to compare (Keeling 2010). The chromerid plastids share features that are retained in either apicomplexan (four membranes, conserved gene order) or dinoflagellate plastids (form II Rubisco acquired by horizontal gene transfer, transcript polyuridylylation, thylakoids stacked in triplets) and encode a full complement of their reduced gene sets (Janouškovec et al. 2010). Finally, with the ability to make plastid genome phylogenies, the chromerids provided concrete evidence that the extant plastids of apicomplexans and dinoflagellates were inherited from a common red algal endosymbiont (Janouškovec et al. 2010).

**The discovery of apicomplexan-like lineages (ARLs)**

The discovery of the chromerids spurred an investigation into the existence of other photosynthetic relatives of the apicomplexans. Interestingly, the identification of eukaryotic plastid contamination in environmental surveys aimed at bacteria (16S rRNA meta-barcoding) revealed plastid sequences related to apicomplexan plastid homologues, which were globally distributed and strongly associated with corals (Janouškovec et al. 2012). These Apicomplexan-Related Lineages (ARLs) include sequences related to *Chromera* and *Vitrella* but also include several abundant groups with no characterized members (Figure 3). In particular, ARL-V – the most abundant – is closely related to the parasitic apicomplexans and has been inferred to be a potential photosynthetic coral symbiont, similar to zooxanthellae (*Symbiodinium*), as it was found enriched in healthy coral tissue and shallow reef depths (Janouškovec et al. 2013).

Traditionally, coral has not been considered a significant host for apicomplexans: only one species of coral-infecting apicomplexan has been formally described based on morphology alone, *Gemmocystis cylindrus* (Upton and Peters 1986a), and one clade of environmental sequences has been described as coral-associated, called ‘Genotype- N’ (Toller et al. 2002). Kirk
et al. showed some prevalence, seasonal differences and potential transmission strategies of ‘Genotype-N’ (Kirk, Ritson-Williams, et al. 2013; Kirk, Thornhill, et al. 2013). Therefore, the discovery of the chromerids and these enigmatic ARL lineages, both which are associated with coral reefs, raise promising questions about the role that coral might have played in the early evolution of apicomplexans.

Figure 3: Current state of molecular data available for apicomplexans and chrompodellids. Groups or taxa shown in purple have genomes available for both the plastid and nucleus. Taxa shown in red have transcriptomic (nuclear) data available. ARL lineages shown in blue only have the plastid 16S SSU rRNA gene sequenced. Groups shown in yellow have nuclear 18S SSU rRNA gene sequences. * indicates the presence of two draft nuclear genomes (Ascogregarina taiwanensis and Gregarina niphandrodes) available in the eugregarines, however the majority of the group only has the 18S SSU gene available. Circles indicate a close association with coral.
1.5 Research Objectives

Genomic methods have already transformed our understanding of the evolutionary origins of apicomplexans and their cellular and ecological contexts, however this picture remains incomplete. Most critically, we lack a well-resolved apicomplexan phylogeny that contains representatives from all major apicomplexan groups, or comprehensive data on the distribution of the relatively well-studied pathways and features that define the group. More specifically, we currently know a great deal about the vertebrate-infecting apicomplexans in the three well-studied subgroups, coccidians, piroplasms, and haemosporidians, and increasingly about the free-living relatives of apicomplexans, but comparatively little about most of the apicomplexan lineages that branch between the two. By exploiting single-cell transcriptomics methods (Picelli et al. 2014) that eliminate the need of a culture or large numbers of isolates, I aim to sample a broad diversity of invertebrate-infecting apicomplexans for high-throughput sequencing surveys (Figure 4). The central goal of these aims is to gain insights into the origin and early evolution of apicomplexans, with a focus on reductive organelle evolution by examining the diversity of apicoplasts and mitochondria in deep-branching apicomplexans. I will generate a robust and taxon-rich apicomplexan phylogeny which will further our understanding of the origins and evolutionary history of the unique characters in this group, and clarify their role in the transition from free-living ancestors to obligate parasites. Specific goals include:

1. Examine the diversity and distribution of the chromerids and ARLs in large-scale environmental surveys to refine and substantiate the relationship of the chromerids and ARLs with coral. This will provide insights into the role of coral hosts in apicomplexan evolution.

2. Generate a robust apicomplexan phylogeny with the use of concatenated nuclear markers by generating transcriptomic and genomic datasets from understudied apicomplexan groups such as the gregarines, and other invertebrate-infecting taxa. This will provide a well-resolved and reliable taxonomic review of the Apicomplexa.

3. Reconstruct the presence and function of apicoplasts across the diversity of apicomplexans by carrying out apicoplast genome sequencing in groups that are missing plastid data.
4. Generate high-quality transcriptomes from deep-branching apicomplexan taxa to bioinformatically reconstruct the evolution of apicomplexan cellular structures and metabolism (ties into Objective 2).

**Figure 4:** Maximum likelihood (ML) 18S rRNA gene phylogeny of the gregarines. Clades that are coloured black indicate gregarine groups that nuclear data is available for. Taxa that are labelled in blue denote target species for transcriptome sequencing (Objective 2 and 4). Black dots on nodes indicate >90% support ($n = 1,000$ ultrafast bootstrap replicates).
Chapter 2: Global diversity and distribution of close relatives of the apicomplexan parasites

2.1 Introduction

Although apicomplexans are obligate intracellular parasites of animals, they have been found to harbour a non-photosynthetic plastid, known as an apicoplast (McFadden et al. 1996). This discovery was particularly intriguing as it suggested that apicomplexan parasites evolved from free-living, photosynthetic ancestors. Our understanding of this transition was greatly aided by the subsequent discovery of Chromera velia, a photosynthetic relative of apicomplexans that was isolated from the stony coral, Plesiastrea versipora, in Sydney harbour (Moore et al. 2008). A second photosynthetic relative of apicomplexans, Vitrella brassicaformis, was also discovered from the stony coral, Leptastrea purpurea, from the Great Barrier Reef (Oborník et al. 2012). Together, Chromera and Vitrella are often referred to as ‘chromerids’, a non-monophyletic group that belongs to a larger clade (chrompodellids) that includes non-photosynthetic predators called colpodellids (Janouškovec et al. 2015).

The ecology of the chromerids remains surprisingly speculative, given they are the first new group of algae to be described in almost 100 years. As they were both isolated from coral, they are widely assumed to be coral symbionts (Moore et al. 2008; Okamoto and Mcfadden 2008; Oborník et al. 2012) however this has not actually been shown in nature. The genomes, evolutionary history and biochemical pathways of chromerids have been extensively studied (Janouškovec et al. 2010; Gile and Slamovits 2014; Flegontov et al. 2015; Woo et al. 2015). In contrast, however, their complete sexual life cycle, whether they in fact live within coral, and if so their role within the coral holobiont, have all yet to be determined beyond a few preliminary observations. Cumbo and colleagues (2013) showed that C. velia can be transmitted vertically from adult Montipora digitata to the coral’s egg (Cumbo et al. 2013). And Mohamed et al. (2018) showed that the coral’s transcriptomic response to Chromera might suggest a parasitic relationship to coral (Mohamed et al. 2018).
These questions became even more complex with the first molecular analyses of their
distribution in nature. Traditionally, coral reefs have not been considered a significant habitat for
apicomplexans: only one species of coral-infecting apicomplexan has been formally described
based on morphology alone, *Gemmocystis cylindrus* (Upton and Peters 1986a), and one clade of
environmental sequences has been described as coral-associated, called ‘Genotype- N’ (Toller et
al. 2002). Kirk et al. showed some prevalence, seasonal differences and potential transmission
However, identifying eukaryotic plastid contamination in environmental surveys aimed at
bacteria revealed that there is actually a large diversity of plastid sequences related to
apicomplexan plastid homologues, which are globally distributed and strongly associated with
corals (Janouškovec et al. 2012). These Apicomplexan-Related Lineages (ARLs) include
sequences related to cultured chromerids, but also include several abundant groups with no
classified members. In particular, ARL-V – the most abundant – is closely related to the
parasitic apicomplexans and has been inferred to be a potential photosynthetic coral symbiont
(Janouškovec et al. 2013).

The ability to detect plastid rRNA genes in bacterial surveys opens the door to detailed analyses
of how eukaryotic algae are distributed in nature (del Campo et al. 2017). However, the initial
surveys that identified ARL-V were limited in scope because of their use of a few available full-
length environmental sequences as opposed to the short-read database, which has exploded in
size since the identification of ARLs. While some association between ARL-V and coral was
shown in their study, there were insufficient data for any solid conclusions about the distribution
of either of the characterized chromerids (Janouškovec et al. 2013). Here, we take advantage of
the now-extensive short read survey data to curate a more comprehensive dataset based on 50
bacterial rRNA amplicon surveys comprising 220 million sequences from 224 unique
geographical locations. This data emphasizes surveys not only from several environments
associated with coral but also includes a wide variety of non-coral environments to test the
association of ARLs both with coral environments versus marine environments more generally,
and with the coral host versus the coral reef more generally.
2.2 Results and Discussion

We retrieved 219,103,555 bacterial SSU rRNA sequences from 50 high-throughput amplicon sequencing surveys from the NCBI SRA (Table A1). From this dataset, we recovered 94,324 plastid SSU rRNA sequences representing 4072 OTUs that represent apicomplexan-related lineages (ARLs). By far the most abundant and diverse lineage retrieved was ARL-V. This lineage was comprised of 92,039 sequences and 3,835 OTUs that formed eight distinct clades, four of which were previously unrecognized (Figure 5). The next most abundant group was Vitrella and its close relatives, where 1860 sequences and 210 OTUs forming four distinct clades were recovered. Other ARLs were in much lower abundance (Figure 5), including Chromera, where only 17 reads representing 7 OTUs were identified and ARL-II with 259 reads and 5 OTUs. In addition, two new lineages were also identified, ARL-X and ARL-XI, with ARL-X branching as a close sister to Vitrella (Figure 5).

Distribution of ARLs confirms association with coral reefs

It has been proposed that ARLs are specifically coral reef associated (Janouškovec et al. 2013). However, these conclusions were based on only 121 ARL sequences. We tested this association with this larger dataset and with numerous environmental outgroups (70% of sequences in our dataset were from coral reef studies, and 30% consist of surveys of other marine and freshwater environments). Our findings confirm that the link between the ARLs and coral reefs is robust. We analysed samples from 224 unique geographical sampling sites worldwide and ARLs were found in only 29 of these locations, all which correspond to coral reefs (Figure A1). One exceptional result was the presence of ARL-V reads in landlocked lakes in central Asia (Baatar et al. 2016). However, we noted that coral bacterial communities were being analysed by the same authors at a similar time, and we conclude that these low abundance reads most likely represent cross-contaminants (Lee et al. 2016).
Figure 5: Maximum likelihood phylogenetic tree inferred from the plastid 16S rRNA gene. The coloured clades show the recovered genes of Apicomplexan-Related Lineages (ARLs) in this study in relation to core apicomplexans, and a dinoflagellate outgroup. The non-colored taxa are the phylogenetic framework from a reference phylogeny. For each plastid lineage, the total number of reads recovered are shown in parentheses. The tree was constructed using RAxML Evolutionary Placement Algorithm with a reference framework of known 16S rRNA genes of apicomplexans and ARLs. Black circles correspond to nodes with bootstrap support greater than 70%.
**Chromera and Vitrella are not coral symbionts**

As its isolation from coral using methods used to extract *Symbiodinium*, *Chromera* has been widely implied and assumed to be a coral symbiont (Moore et al. 2008; Cumbo et al. 2013). However, this has never been shown directly and, according to a recent publication, *Chromera* is more likely an opportunistic parasite than a mutualistic symbiont (Mohamed et al. 2018). Here, we show that *Chromera* is relatively rare in coral reefs worldwide, and, interestingly, is also never recovered from the actual coral: *Chromera* reads were only detected in biogenous sediments surrounding corals, and never from coral tissue itself (Figure 6). A possible limitation is the absence of survey data from the coral species where Chromera was originally isolated, *Plesiastrea versipora* and *Leptastrea purpurea*. However, given how strongly ARL-V sequences are associated with specifically coral mucus and tissue (see below), it seems likely that if *Chromera* were indeed a symbiont, a similarly tight association to the coral host should be detected. Therefore, we suggest that *Chromera* is not a coral symbiont per se. Instead, *Chromera* may be a symbiont of other reef-dwellers that have yet to be sampled, or a free-living autotroph in the reef biogenous sediments. Its association with the reef environment could be due to the distribution of another invertebrate host, or perhaps it's life cycle includes a transient stage within coral that has not been observed but ties it to the reef environment. Sediments are known to harbour a diverse array of primary producers, including cyanobacteria and photosynthetic protists (e.g., dinoflagellates and diatoms) and have been reported to have a largely uncharacterized bacterial SSU diversity (Werner et al. 2008). However, their role in the reef ecosystem is still comparatively understudied. Greater observation of *Chromera* in nature is needed to understand its basic biology, which may help illuminate the ecological conditions driving the evolution of apicomplexans. *Chromera* may well be tied to the microbial processes in biogenous sediments and their contribution to reef primary production, nutrient cycling and maintenance of overall coral reef health.

Interestingly, *Vitrella*-related groups are not only more widely distributed, as sequences were recovered predominantly from the biogenous sediments, like *Chromera*, but also found at lower abundances in coral tissue, mucus and in coral-associated seawater (Figure 6). *Vitrella* has been shown to have both photosynthetic and predatory stages, as well as a large sporangium stage that
has not been seen in *Chromera* (Oborník et al. 2012). The more varied distribution we observe might reflect a more complex life cycle with different trophic stages that take advantage of different parts of the coral reef. Once again, this emphasizes the importance of elucidating the complete life cycle of *Vitrella* in nature to provide some insights into its roles on the reef, and also in the evolution of parasitism in apicomplexans.

**Figure 6**: The proportion of Apicomplexan-Related Lineages (ARLs) recovered from specific coral reef environments.

The proportion has been normalized to the total number of reads retrieved from that environment. Individual clades of each ARL lineage are shown as separate bar graph bars. Shades of red indicate coral tissue and coral mucus. Blue indicates coral-associated seawater and yellow indicates biogenous sediments. Outliers were excluded from this bar graph, refer to Figure A3 for the outlier analysis.
Refining ARL-V and its association with coral

Apicomplexan Related Lineage-V is the most abundant ARL, but remains an environmental clade described solely from plastid SSU gene sequences. Based on the initial 121 environmental sequences, phylogenetic analyses placed ARL-V in an intermediate position between the photosynthetic chromerids and parasitic apicomplexans and showed a very tight association with coral reefs (Janouškovec et al. 2012). A follow-up survey showed ARL-V to be enriched in healthy coral tissue in shallow reefs, suggestive of a photosynthetic symbiotic relationship with coral (Janouškovec et al. 2013). Here, we retrieved 92,039 sequences of ARL-V, which allows us to test inferences about the distribution and ecology of this enigmatic group with greater confidence. First, the new dataset clearly confirms that ARL-V is indeed exclusively coral reef-associated and globally distributed in reefs worldwide. The relative abundance of ARL-V compared with all other ARLs is also clearly confirmed: ARL-V sequences were more than 50-fold more abundant than any other ARL sequences. Lastly, and in striking contrast to chromerids, we confirmed that ARL-V is strongly associated with the coral host directly (in tissue and mucus); it is completely absent in the sediments, and only a few sequences were found in seawater surrounding coral (Figure 7). In addition to confirming its abundance, the larger dataset has also confirmed and extended the high level of diversity of ARL-V. We not only identified eight distinct phylogenetic subgroups of ARL-V but also 3835 OTUs97. Altogether, ARL-V represents several distinct, species-rich lineages, suggesting it may have a high degree of biological diversity as well.

To examine the host range of ARL-V, we searched sequencing surveys of 22 distinct coral species, out of which ARLs were present in 19 species (Figure 7). ARL-V was the most widespread ARL retrieved and was found in every coral host that was positive for any ARL. In addition, ARL-V was retrieved in highest abundance compared with any other ARL in all individual coral species, except for Diploria strigosa (Figure 7). Interestingly, the coral species positive for ARL-V do not form a monophyletic group, but rather included a diverse species of both hexacorals and octacorals (soft corals). In contrast to Janouškovec and colleagues (2012), we also found that ARL-V is not depth dependent: ARL-V was retrieved from corals at depths of 522m as well as in shallow corals at 1m, as well as a range of depths between these extremes.
Moreover, ARL-V was found in four species of deep, cold-water azooxanthellate corals, indicating no obligatory co-occurrence between ARL-V and *Symbiodinium*.

**Figure 7:** The absolute abundances of the Apicomplexan-Related Lineages (ARL) sequences recovered in 24 coral species. Cladogram (left) indicates relatedness of coral species (adapted from (Fukami et al. 2008; Kayal et al. 2013)). Table (right) shows the number of coral samples and corresponding number of reads that were retrieved from those samples. The bar graph shows absolute abundances of ARL reads. Shades of blue indicate ARL-V clades, from ARLV-1 (darkest) to ARLV-8 (lightest). Shades of red indicate *Vitrella* clades, from Vit-1 (darkest) to Vit-4 (lightest). Brown indicates ARL-IV. Clear bar graph bars indicate that no ARLs reads retrieved from those species.
Based on these findings, we propose that the distribution of ARL-V is more consistent with an intracellular parasite or commensal, rather than a photosynthetic symbiont, or that at least some fraction of ARL-V diversity is non-photosynthetic. Incomplete metadata did not allow any comprehensive correlation with coral health, but ARL-V was retrieved from samples of both healthy and diseased coral tissue, and ARL-V has previously been found in healthy corals (Janouškovec et al. 2013). This suggests that ARL-V collectively does not cause any one distinctive coral disease. The small number of sequences found in seawater surrounding coral (Figure 6) might indicate the presence of infective stages, however, it is also possible these are dead or dying cells released from coral, and direct observation of ARL-V and its biology will be needed to elucidate the exact nature of its interaction with coral. This may be problematic given the diversity of ARL-V (Figure 5), which might reflect a high degree of host-specificity or complex population structure within each coral host species, the details of which will require careful sampling of many individual coral hosts.

Overall, ARL-V is the most abundant apicomplexan-related lineage, one that is both phylogenetically and ecologically distinct from the chromerids. While initially speculated to be a photosynthetic symbiont, here we show its distribution to be more indicative of a non-photosynthetic coral parasite or commensal. This bolsters the idea that the ARL-V clade might correspond to the coral apicomplexan Genotype-N (Toller et al. 2002). Genotype-N was discovered based on the nuclear 18S SSU gene, whereas the ARL-V is based on plastid 16S SSU gene. This incompatibility in molecular data prevents us from making any conclusive decision as to whether these organisms are the same (in particular as relatively few nuclear SSU rRNA surveys of coral reefs have been conducted). Another possibility is that ARL-V corresponds to *Gemmocystis cylindrus*, a coccidian has only been identified by morphology (Upton and Peters 1986a). However, once again the data are impossible to compare as neither ARL-V nor Genotype-N cells have knowingly been observed. The relationship between these three entities is a question that remains to be tested directly.
2.3 Chapter 2 conclusions

The association between coral and its archetypal dinoflagellate symbiont, *Symbiodinium*, is very well-studied (Yellowlees et al. 2008; Thornhill et al. 2017), yet relatively little is known about the interactions between coral and other microbial eukaryotes (Ainsworth et al. 2017; del Campo et al. 2017). The *Symbiodinium* model looms large in our thinking about coral-protist interactions, so it is perhaps it is not surprising that chromerids were assumed to be photosynthetic symbionts of a similar sort when first discovered. This also might have played a role in the surprisingly sparse attention paid to their ecological role in nature compared with their biochemistry and evolution. Our analyses suggest, however, that these interactions are both complex, not what we expected, and in need of direct observations in nature. *Chromera* and *Vitrella* are indeed coral reef associated but live within the biogenous sediments and not with coral itself, suggesting a role in reef sediment primary production and some unclear link to the reef ecosystem that is subtler than simply living inside coral cells. We also confirmed ARL-V to be the most abundant, phylogenetically diverse ARL lineage and to associate almost exclusively with coral mucus and tissue of a very broad range of coral hosts. However, the environmental distribution of ARL-V is not consistent with a photosynthesis-based symbiosis either, suggesting instead parasitism or commensalism with coral that also requires direct observation. These organisms have attracted a great deal of attention because they are related to an important and evolutionarily interesting group of parasites, but evidence is mounting that the ecology of apicomplexan-related lineages is not what we anticipated and observing them in nature with specific data on their interactions with coral is a next critical step to understanding both these organisms and their contribution to coral reefs.
2.4 Materials and methods

Amplicon sequence retrieval


Selecting and clustering plastid 16S rRNA gene sequences

Putative apicomplexan related sequences were identified using BLASTn (Camacho et al. 2009) with a BLAST database constructed from the aligned environmental sequences from Janouškovec and colleagues (2012) in addition to publicly available plastid SSU rRNA sequences from apicomplexans and chromerids. Sequences were selected on the criteria that they had a minimum 85% sequence similarity to apicomplexan-related lineages (ARLs) and that the length of the alignment covered at least 90% of the length of the query sequence. Selected sequences were then clustered to a 97% threshold using CD-HIT (Li and Godzik 2006).
Building reference phylogeny and alignment

SSU rRNA sequences of all known ARLs (including *V. brassicaformis* and *C. velia*) were extracted from GenBank and clustered at 97% identity using USEARCH v7.0.1090 (Edgar 2010). Sequences were aligned using MAFFT auto mode (Katoh and Standley 2013) and a set of representative sequences as outgroups (apicomplexans, dinoflagellates and colpodellids). Alignments were checked using AliView (Larsson 2014) and highly variable regions of the alignment were removed using trimAl 1.2 (settings: -gt 0.3 -st 0.001) (Capella-Gutierrez et al. 2009). Maximum likelihood (ML) phylogenetic trees were constructed with RAxML v8 (GTR-CAT-I substitution model, 1000 independent tree searches starting from distinct random topology, and 1000 standard non-parametric bootstraps) (Stamatakis 2014). The reference tree is shown in Figure A2.

Short reads assignment and EPA tree reconstruction using QIIME

QIIME (Quantitative Insights into Microbial Ecology) v1.4.0 (Caporaso et al. 2010) was used to construct a phylogenetic tree to curate the selected reads and discard those that branched within apicomplexans and dinoflagellates. This was done using the following QIIME scripts: (i) align_seq.py, the BLAST selected sequences and the reference alignment were then merged and aligned and (ii) filter_alignment.py, the merged sequences were filtered (a similar process to trimming) (settings: filter_alignment.py, -g 0.99 -s -e 0.0001). RAxML Evolutionary Placement Algorithm (EPA) (model GTR-CAT-I) was used to place the sequences onto the fixed topology of the reference tree (Stamatakis 2014). Long branches and sequences that branched within apicomplexans and dinoflagellates were discarded. This process was carried out until only reads that branched within the apicomplexan-related lineages (ARLs) remained, and these curated sequences were then used for all subsequent analyses.

QIIME analysis: operational taxonomic unit (OTU) picking

VSEARCH v2.4.2 (Rognes et al. 2016) was used to quality filter and remove any chimeric reads. The sequences were then clustered to 97% identity using USEARCH v8.1.0 (Edgar 2010). To assign taxonomy to the reads, we used QIIME’s OTU (Operational Taxonomic Unit) picking pipeline (Caporaso et al. 2010). OTUs were first picked based on sequence similarity within the
reads using pick_otus.py. As each OTU was made up of many related sequences, a representative sequence from each OTU was then picked using pick_rep_set.py. These representative sequences were used for taxonomic annotation of the OTU and phylogenetic alignment in downstream analyses. We assigned taxonomy to each OTU representative sequence with assign_taxonomy.py, using the UCLUST consensus taxonomy classifier (Edgar 2010). Using this annotation, a table of OTU abundances in each sample with taxonomic identifiers for each OTU was constructed. OTUs were further filtered according to their observation counts and OTUs that were observed fewer than two times (i.e., singletons) were discarded. The final phylogenetic tree was constructed using these curated sequences. They were assigned positions on to the fixed topology of the reference tree using the EPA algorithm implemented in RAxML (Stamatakis, 2006) assuming the GTR-CAT-I substitution model. New clades were assigned if reads did not branch with any of the reference taxa. Tree figures were edited with FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/) (Figure 5).

**Ecological meta-analyses**

A metadata table was collated by manually downloading all the ecological information associated with the curated sequences, such as geographical location, environmental material, host, etc., from the NCBI SRA (Leinonen et al., 2011). Using the OTU taxonomy and the associated metadata, QIIME summarize_taxa_through_plots.py was used to identify patterns and differences in the relative abundance of the recovered ARLs in different coral hosts and environments (Figures 6 and 7).

### 2.5 Chapter-specific acknowledgments

We thank Racquel Singh and Nick Irwin for their valuable help on the data analyses. This work was supported by a grant from the Canadian Institute for Health Research to PJK (MOP 42517). JdC and MK were supported by a grant to the Centre for Microbial Diversity and Evolution from the Tula Foundation and JdC was supported by a Marie Curie Outgoing Fellowship grant (FP7-PEOPLE-2012-IOF – 331450 CAARL)
Chapter 3: Multiple independent origins of apicomplexan-like parasites

3.1 Introduction

Apicomplexans are an extremely diverse and specious group but are nevertheless united by a distinctive suite of cytoskeletal and secretory structures related to infection, called the apical complex, which is used to recognize and gain entry into animal host cells (Votýpka et al. 2017). The apicomplexans are also known to have evolved from free-living photosynthetic ancestors and retain a relict plastid (the apicoplast), which is non-photosynthetic but houses a number of other essential metabolic pathways (Lim and McFadden 2010). Genomic analyses of their free-living relatives (chromerids and colpodellids) have revealed a great deal about how the alga-parasite transition may have taken place, as well as origins of parasitism more generally. Here, we show that, despite the surprisingly complex origin of apicomplexans from algae, this transition actually occurred at least three times independently. Using single-cell genomics and transcriptomics from diverse uncultivated parasites, we find that two genera previously classified within the Apicomplexa, Piridium and Platyproteum, form separately branching lineages in phylogenomic analyses. Both retain cryptic plastids with genomic and metabolic features convergent with apicomplexans. These findings suggest a predilection in this lineage for both the convergent loss of photosynthesis and transition to parasitism, resulting in multiple lineages of superficially similar animal parasites.

3.2 Results and Discussion

To gain a deeper understanding of the origin of parasitism in apicomplexans, we used single-cell sequencing to characterize the genomes and transcriptomes from a number of uncultivated parasites representing poorly studied lineages of apicomplexans. Specifically, we generated transcriptome data from individual trophozoite cells of the gregarine apicomplexans Monocystis agilis, Lecudina tuzetae, Pterospora schizosoma, Heliospora capraellae, and Platyproteum sp., using single cells documented microscopically and manually isolated directly from their animal hosts (Figures 8A–E). In addition, we generated both genomic and transcriptomic data from
gamogonic stages of *Piridium sociabile*, an apicomplexan isolated from the foot tissue cells of the common marine whelk, *Buccinum undatum* (Figure 8F). These gregarines represent subgroups of both marine (*Pterospora*, *Heliospora*, *Lecudina*, and *Platyproteum*) and terrestrial (*Monocystis*) parasites, and the limited available molecular data (from small subunit [SSU] rRNA) are divergent but generally show them to be diverse, early branching apicomplexans (Leander, Clopton, et al. 2003; Leander et al. 2006; Rueckert, Villette, et al. 2011; Rueckert, Simdyanov, et al. 2011) (Figure 8B). *Platyproteum* was the most recently described by detailed microscopy and molecular phylogenetic analyses using SSU rDNA sequences; these data suggest that it is a particularly deep-branching apicomplexan (Leander 2006; Rueckert and Leander 2009). *Piridium sociabile* is even more poorly studied: found in 1932 as an intracellular infection and was morphologically classified as a schizogregarine (Patten 1936).

The relationships of these six taxa to the Apicomplexa were examined by phylogenomics using a concatenated alignment of 39 taxa and 189 nucleus-encoded proteins that have been previously used in in both eukaryote-wide and phylum-level phylogenies (Burki et al. 2016; Irwin et al. 2019). Their positions in the resulting tree are strongly and consistently resolved by both maximum likelihood (C40+LG+G4+F model) and Bayesian (CAT-GTR) analyses (Figure 8G). Surprisingly, the phylogeny shows that neither *Piridium* nor *Platyproteum* branch within the Apicomplexa. Instead, *Piridium* branches within the sister group to the Apicomplexa, the chrompodellids, with complete support as sister to the photosynthetic alga *Vitrella brassicaformis*. *Platyproteum* forms a new lineage, also with complete support, sister to the clade consisting of apicomplexans and chrompodellids collectively. The four more canonical gregarines (*Monocystis*, *Lecudina*, *Pterospora*, and *Heliospora*) formed a monophyletic group of deep-branching apicomplexans that interestingly excludes *Cryptosporidium*. This robust phylogeny not only confirms that photosynthesis was lost multiple times independently around the origin of the Apicomplexa but more surprisingly shows that the highly derived mode of animal parasitism that is characteristic of the Apicomplexa also arose multiple times independently.
Figure 8: Phylogenomic tree of the Apicomplexa and relatives.

Light micrographs of single-cell trophozoites are of (A) H. caprellae, (B) L. tuzatae, (C) M. agilis, (D) P. schizosoma, and (E) Platyproteum sp. (scale bars represent 50 µm). (F) Light micrograph of a single-cell gamont of P. sociabile (scale bar represents 15 µm). (G) Maximum likelihood tree generated from an alignment comprising 198 genes and 58,116 sites under the C40+LG+G4+F substitution model with both non-parametric bootstraps (n = 500) and posterior probabilities (PPs) shown. Black circles represent 100% bootstrap support and 1.0 Bayesian PP, and all other support values are indicated beside the node. New transcriptomes are shown in bold lettering. The percentage of genes present in the phylogeny for each taxon is shown on the left and is shaded in black for newly sequenced transcriptomes. On the right are characters corresponding to each taxon.

To further investigate the convergent evolution of parasitic lifestyles in *Piridium* and *Platyproteum*, we examined plastid retention and function, a well-studied trait of the Apicomplexa (Janouškovec et al. 2015; McFadden and Yeh 2017). With both genomic and transcriptomic data from *Piridium*, we first assembled its complete plastid genome (Figures 9A and 9B), which is strikingly similar in size, architecture, and gene content to apicoplast genomes (Figure 9B). The *Piridium* plastid genome is a highly reduced compact circle (~34 kbp) with all
remaining genes in perfect synteny with homologs in its closest relative, the photosynthetic *Vitrella*. Similar to the apicoplast, it is extremely AT rich (21% G+C content) and uses a non-canonical genetic code where UGA encodes tryptophan (as seen in *Chromera*, *Toxoplasma*, and corallicolids, but not in the more closely related *Vitrella*) (Janouškovec et al. 2010; Kwong et al. 2019). It retains similar ribosomal genes as well as the same bacterial RNA polymerases (*rpoB*, *rpoC1*, and *rpoC2*) and other protein-coding genes (*sufA*, *clpC*, and *tufA*) as apicoplasts. It has also convergently lost all genes relating to photosynthesis, as well as *rps18, rpl13, rpl27, secA*, and *secY* (Figure 9C). Reflecting its origin from a chrompodellid ancestor, the *Piridium* plastid also encodes a handful of genes that are present in *Vitrella* but absent from apicoplasts: *rps14; rpl3; and rpoA*. Curiously, only a partial rRNA inverted repeat remains in *Piridium*; this repeat is ancestral to all apicomplexans and chrompodellids but has also similarly been lost in the piroplasm apicomplexans, *Babesia* and *Theileria* (Gardner et al. 2005; Huang et al. 2015).

Apicomplexans depend on apicoplasts for essential biosynthesis of four compounds: isoprenoids (using the non-mevalonate pathway); heme; iron-sulfur (Fe-S) clusters; and fatty acids (type II fatty acid pathway) (McFadden and Yeh 2017). All apicomplexans rely on these pathways except piroplasms, which have lost the FASII and heme pathway and use cytosolic FASI instead, and *Cryptosporidium*, which can salvage the metabolites from its host and has lost its plastid entirely(Abrahamsen et al. 2004; Brayton et al. 2007). We identified all enzymes from these pathways and all enzymes for analogous and homologous cytosolic pathways using profile hidden Markov models (HMMs) and analyzed the resulting genes for evidence of distinctive N-terminal bipartite plastid-targeting peptides (Figure 9C; Table B1). It is impossible to conclude that any single gene is absent based on transcriptomic data alone, so only the absence of all genes for entire biochemical pathways is considered here. The dependency on plastid metabolism in *Piridium* is identical to most apicomplexans, with the retention of all four pathways but no photosystems or other known plastid functions. *Platyproteum* is similar but has also lost the FASII pathway and so more resembles the piroplasms (Gardner et al. 2005; Brayton et al. 2007).
Figure 9: Plastid dependency in *Piridium* and *Platyproteum* has evolved convergently to apicomplexans.

(A) Complete annotated plastid genome of *P. sociabile*. (B) Presence of plastid biosynthetic pathways across the tree of apicomplexans and chrompodellids. Portions of the circles represent the proteins found in each pathway found (key shown on right). Black circles indicate the presence of complete N-terminal bipartite plastid-targeting peptides (only shown for newly added transcriptomes). (C) Plastid gene content of apicomplexans and *Vitrella* (free-living, photosynthetic) compared to *Piridium*. 
Interestingly, the same analysis on the clade of gregarines revealed a greater degree of variation from other apicomplexans than seen in the cryptic plastids that evolved in parallel (Figure 2B). Like Cryptosporidium, the terrestrial gregarines Monocystis and Gregarina have completely lost all plastid metabolism and likely also lost the organelle (which also suggests that the phylogenetic relationship between Cryptosporidium and terrestrial gregarines remains uncertain) (Abrahamsen et al. 2004; Toso and Omoto 2007). In contrast, however, the marine gregarines Lecudina and Pterospora retain the complete FASII pathway but no other identifiable plastid metabolism. This is the first evidence of a plastid in any gregarine and is also functionally curious, because it is isoprenoid biosynthesis that has been proposed to be the main barrier to plastid loss (Janouškovec et al. 2015). The gregarines thus suggest that plastid dependency is highly context dependent.

Looking beyond the plastid, metabolic reconstructions based on KEGG (Kyoto Encyclopedia of Genes and Genomes) identifiers across the whole genome confirm an overall convergence of functional reduction but also some divergence (Figure 10). Both Piridium and Platyproteum have, as expected, substantially reduced their metabolic functions compared to their free-living chrompodellid relatives. However, neither is as reduced as apicomplexan parasites. In both cases, a few core pathways, such as the glyoxylate cycle and pyrimidine catabolism, have been retained (Table B2). Of the two, Piridium contains the greatest breadth of biosynthetic functions that were mostly lost in all other parasitic groups, such as de novo amino acid biosynthesis (isoleucine and arginine) and purine biosynthesis (inosine) and degradation. Surprisingly, the gregarine Monocystis agilis has also retained a greater metabolic capacity than other apicomplexans. Although its greater functional capacity relative to other gregarines may be due to better sequencing coverage, the majority of other apicomplexans are reconstructed from whole genomes, suggesting that the baseline metabolic complexity of the group as a whole is greater than was previously thought.
Figure 10: The distribution of cellular metabolic pathways across the tree of apicomplexans and chrompodellids.

The list of metabolic pathways is shown on the right. Yellow represents presence, and shades of blue indicate absence based on genomic data (dark blue) or absence based on transcriptomic data (lighter blue). Our newly sequenced transcriptomes are shown in bold lettering. Estimated gains and losses of genes (orthogroups) are shown on nodes and on the branches leading to each species. The pie charts show the percentage of genome or transcriptome completeness based on BUSCO scores.
3.3 Chapter 3 conclusions

The origin of apicomplexan parasites from free-living photosynthetic alga represents a major evolutionary transition between two very different modes of living, so different in this case that the idea was originally met with skepticism. The current data show that, however dramatic this transition may seem, it was not unique but rather repeated at least three times in related lineages of photosynthetic algae. The details of the parasitic machinery in *Piridium* and *Platyproteum* are unknown, so how detailed the convergence of their parasitic lifestyles may be will require more information, but they superficially resemble apicomplexans to the extent that they were classified within the group when formally described. The genomic and transcriptomic data presented here also suggest that the ancestors of these lineages maintained high levels of redundancy in metabolic pathways between compartments that persisted over long periods of evolutionary time and apparently shared some predilection to animal parasitism. The underlying reason for this is not clear, because the evolution of apicomplexan parasitism is not linked to the acquisition of any novel feature or machinery but is instead marked by loss and tinkering of the existing genomic repertoire.

3.4 Materials and Methods

*Amplion Genomics and transcriptomics of Piridium sociabile*

DNA and RNA from the resulting gamonts was then extracted using a QIAGEN, Allprep DNA/RNA Mini Kit (Cat. No. 80204). cDNA was synthesized using the SMARTseq2 protocol with seven cDNA amplification cycles (Picelli et al. 2014). RNA and DNA sequencing libraries were both prepared using Illumina Nextera XT and Nextera protocol respectively, and sequenced using 2x300bp Illumina MiSeq (DNA) and 2x100bp Illumina HiSeq 2000 run (RNA). Both RNA and DNA reads were adaptor and quality trimmed with Trimomatic (Bolger et al. 2014). RNA reads were further processed to remove low complexity regions using PRINTseq (Schmieder and Edwards 2011) and were assembled into transcripts using Trinity v2.4 (with default settings) and translated into protein sequences using Transdecoder v.5 (Haas et al. 2013).
**Genome assembly and annotation of Piridium sociabile**

The MIRA4 assembler was used to assemble the genomic DNA reads, which led to the assembly of single circular plastid genome chromosome (Chevreux B and Suhai). This assembly was validated by mapping of the reads back to the assembly by Bowtie2 (Langmead et al. 2009). The plastid genome was then automatically annotated using MFAnnot (http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl) and RNAweasel (http://megasun.bch.umontreal.ca/cgi-bin/RNAweasel/RNAweaselInterface.pl), followed by manual corrections in Geneious v11.1.5 (https://www.geneious.com).

**Transcriptomics of the gregarines and Platyproteum**

The single-cell trophozoites were washed at least three times in autoclaved filtered seawater, or ultrapure water (for *Monocystis*) and viewed and photographed under a Leica DMIL LED microscope equipped with a 40x objective and a Sony a6000 camera. Single trophozoite cells were picked using a glass capillary micropipettes and transferred to a 0.2 µL thin-walled PCR tube containing 2 µL of cell lysis buffer (0.2% Triton X-100 and RNase inhibitor (Invitrogen)). cDNA was synthetized from the single cell, or a pool of 2-3 cells, using the Smart-Seq2 protocol (Picelli et al. 2014). The cDNA concentration was quantified on a Qubit 2.0 Fluorometer (Thermo Fisher Scientific Inc.). Prior to high-throughput sequencing, 1L of the final cDNA product was used as a template for a PCR amplification of the V4 region of the 18S rRNA gene using Phusion High-Fidelity DNA Polymerase (New England Biolabs, Thermo Scientific) and the general eukaryotic primer pair TAREuk454FWD1 and TAREukREV3 (Stoeck et al. 2010). The PCR product was then sequenced by Sanger dideoxy sequencing. The SSU rRNA gene sequences were used to confirm the identity of the newly collected organisms and avoid animal host contamination using BLASTn to look for similar sequences in the non-redundant NCBI database (Altschul et al. 1990). Once the identity of the parasite was confirmed, sequencing libraries were prepared using the Nextera XT protocol, and sequenced on a single lane of Illumina MiSeq using 250 bp paired end reads.
Transcriptome assembly and annotation of the gregarines and Platyproteum

The raw Illumina sequencing reads were merged using PEAR v0.9.6, and FastQC was used to assess the quality of the paired reads (Zhang et al. 2014; Andrews et al. 2015). The adaptor and primer sequences were trimmed using Trimmomatic v0.36 and the transcriptomes were assembled with Trinity v2.4.0 (Haas et al. 2013; Bolger et al. 2014). The contigs were then filtered for animal host contaminants using BlobTools in addition to BLASTn and BLASTx searches against the NCBI nt database and the Swiss-Prot database, respectively (Laetsch et al.; Altschul et al. 1990). Coding sequences were predicted using a combination of TransDecoder v3.0.1 and similarity searches against the Swiss-Prot database (Haas et al. 2013). Assessment of the quality of the assembly and annotation of the transcriptomes (including Piridium) was carried out using BUSCO (Simão et al. 2015).

Ortholog identification, gene concatenation and phylogenomics

In addition to our newly generated transcriptomes, the following transcriptomes and genomes were downloaded from EuPathDB and screened for orthologs; Hammondia hammondi, Sarcocystis neurona, Eimeria falciformis, and Gregarina niphandrodes (Aurrecoechea et al. 2017). All transcriptomes were comprehensively searched for a set of 263 genes that have been used in previous phylogenomic analyses (Burki et al. 2016; Irwin et al. 2019). All the sequences in the 263 gene-set, representing a wide range of eukaryotes, were used as queries to search the above datasets using BLASTn (Altschul et al. 1990). The hits were then filtered using an e-value threshold of 1e-20 and a query coverage of 50%. Each of the gene-sets was then aligned using MAFFT L-INS-i v7.222 and trimmed using trimAl v1.2 with a gap-threshold of 80% (Capella-Gutiérrez et al. 2009; Katoh and Standley 2013). Single gene trees were then constructed to identify paralogs and contaminants using IQ-TREE v.1.6.9 (LG+G4 model) or RAxML v8.2.12 (PROTGAMMALG model) with support from 1000 bootstraps (Stamatakis 2014; Nguyen et al. 2015). The resulting trees were manually scanned in FigTree v1.4.2 and contaminants and paralogous sequences were identified and removed (Rambaut 2014). The final cleaned gene-sets were filtered so that they contained only a maximum of 40% missing OTUs and then concatenated in SCaFoS v1.2.5 (Roure et al. 2007). The resulting concatenated alignment consisted of 198 genes spanning 58,116 amino acid positions from 39 taxa. The phylogenomic
maximum likelihood tree was constructed with the heterogenous mixture C40+LG+G4+F model as implemented in IQ-TREE (model LG+G4+F yielded identical topology) (Nguyen et al. 2015). Statistical support was inferred using 500 bootstrap replicates using the LG+C40+G4+F PMSF profiles, and 1000 bootstrap replicates using the LG+G4+F model in RAxML (Stamatakis 2014; Wang et al. 2018). The Bayesian tree was computed using Phylobayes under the GTR-CAT model with constant sites removed from the analyses (Lartillot et al. 2009). Four independent chain were run for 9 thousand generations and converged with maxdiff = 0.19 (20% burning) (Figure 8G). For the plastid based phylogenomic analyses, a previously published dataset (Ševcíková et al. 2015) was used and enriched with proteins from wide sets of publicly available apicoplast proteins and the *Piridium* plastid genome. The plastid phylogenomic tree was constructed using a concatenated alignment of 62 plastid-encoded proteins with RAxML using the LG+G4+F substitution model with 500 bootstrap replicates (Stamatakis 2014) (Figure 8B).

**Search and identification of plastid proteins**

Profile hidden Markov models (HMMs) were used to identify plastid metabolic proteins in our transcriptomes based on curated alignments. To construct the curated alignments, known dinoflagellate proteins were used as queries in a BLASTp search (e-value threshold of 1e-5) against a comprehensive custom database containing representatives comprised of major eukaryotic groups, with a focus on plastid-containing lineages (dinoflagellates, chromodellids, Apicomplexa, cryptophytes, haptophytes, stramenopiles, Archaeplastida) as well as selected taxa from non-plastid lineages (Opisthokonta, Amoebozoa, Apusozoa, Ancyromonadida and ciliates) and RefSeq data from all bacterial phyla at NCBI (https://www.ncbi.nlm.nih.gov/, last accessed December 2017) (Altschul et al. 1990). The database was subjected to CD-HIT with a similarity threshold of 85% to reduce redundant sequences and paralogs (Li and Godzik 2006). Results from blast searches were parsed for hits with a minimum query coverage of 50% and e-values of less than 1e-25 (or 1e-5 for HemD). The number of bacterial hits was restrained to 20 hits per phylum (for FCB group, most classes of Proteobacteria, PVC group, Spirochaetes, Actinobacteria, Cyanobacteria (unranked) and Firmicutes) or 10 per phylum (remaining bacterial phyla) as defined by NCBI taxonomy. Parsed hits were aligned with MAFFT v. 7.212, using the–auto option, poorly aligned regions were eliminated using trimAl v.1.2 with a gap threshold
of 80% (Capella-Gutiérrez et al. 2009; Katoh and Standley 2013). Maximum likelihood tree reconstructions were then performed with FastTree v. 2.1.7 using the default options (Rambaut 2014). The resulting phylogenies and underlying alignments were inspected manually to remove contaminations, recent paralogs and duplicate sequences. The cleaned, unaligned sequences were then subjected to filtering with PREQUAL using the default options to remove nonhomologous residues introduced by poor-quality sequences, followed by alignment with MAFFT GINSi using the VSM option (unalignlevel 0.6) to control over-alignment (Katoh and Standley 2013; Whelan et al. 2018). The alignments were subjected to Divvier (https://github.com/simonwhelan/Divvier) using the divvygap option to improve homology inference before removing ambiguously aligned sites with trimAl v. 1.2 (gap threshold of 1%) (Capella-Gutiérrez et al. 2009). Trees for final sequence curation were calculated with IQ-TREE v. 1.6.5, using the mset option to restrict model selection (to DAYHOFF, DCMUT, JTT, WAG, VT, BLOSUM62, LG, PMB, JTTDCMUT) for ModelFinder, while branch support was assessed with 1000 ultrafast bootstrap replicates, and once more subjected to manual inspection (Nguyen et al. 2015; Hoang et al. 2018).

Profile HMMs were then generated using these curated alignments and HMM searches were conducted on all transcriptomes and genomes using HMMER v3.1 and an e-value threshold of 1e-5 (Finn et al. 2011). All the hits were then extracted and incorporated into the original alignments and realigned using MAFFT v7.222 (–auto option). The resulting alignments were then used to generate phylogenies in IQ-TREE v.1.6.9 using the LG+F+G4 substitution model and statistical support was assessed using 1000 ultrafast bootstrap replicates (Katoh and Standley 2013; Nguyen et al. 2015). The phylogenies were then manually scanned in FigTree v1.4.2 and contaminants, paralogs, mitochondrial sequences, and long-branching divergent sequences were identified and removed. The remaining sequences were realigned and used to generate maximum likelihood phylogenies in IQ-TREE v.1.6.9. Phylogenetic models were selected for each tree individually based on Bayesian Information Criteria using ModelFinder as implemented in IQ-TREE, and statistical support was assessed using 1000 ultrafast bootstrap pseudoreplicates (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017).
Search for plastid localization signals

To investigate the N-terminal extensions and thus intracellular location of proteins of interest, corresponding alignments were manually inspected for completeness of the sequences and for N-terminal extensions relative to prokaryotic or cytosolic homologs. Prediction of signal peptides as part of N-terminal bipartite leader sequences was performed with the Hidden Markov Model of SignalP3.0 using the default truncation setting of 70 residues (Dyrløv Bendtsen et al. 2004). To predict putative N-terminal transmembrane domains, TMHMM v. 2.0 was used only on the first 100 amino acid residues of the transcript to improve prediction accuracy (Sonnhammer et al. 1998). Putative plastid transit peptides were interpreted as 24-aa stretches downstream of the signal peptide, representing the minimum length for apicomplexan transit peptides still within the N-terminal extension and upstream of the estimated start of the mature protein, as described by Parsons et al. (Parsons et al. 2007) Conserved domains and their coordinates in the mature protein region of candidate sequences were identified with the Pfam sequence search service on http://pfam.xfam.org/search/sequence, using the gathering threshold as a cut-off (Table S1).

Analysis of cellular metabolic pathways

We reconstructed the metabolic maps for our new transcriptomes, as well as representative species across the apicomplexan and chrompodellids, using Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2016). We first assigned KEGG ortholog identifiers (KO) to all proteomes using the web-based server, KAAS (KEGG Automatic Annotation Server) (https://www.genome.jp/kegg/kaas/), and where possible we used annotations already available within KEGG. The assigned KO numbers were used to identify complete metabolic pathways using the KEGG reconstruct module and module mapper. Complete metabolic pathways present in Piridium, Platyproteum or both but missing in other apicomplexans were further investigated. The identity of all proteins in these unique pathways were confirmed using BLASTp to ensure removal of contaminants and false positives (Table B2).
3.5 Chapter-specific acknowledgments

We thank Maria Herranz, Niels van Steenkiste, Phil Angel, and the Hakai Institute for their assistance in sample collection. We also thank Filip Husnik for his guidance with the transcriptomics and Waldan Kwong for his valuable feedback on the manuscript. This work was supported by grants from the Canadian Institutes for Health Research (MOP-42517) to P.J.K. and HIR grant (UMC/ 625/1/HIR/027) to M.A.F. from the University of Malaya, Kuala Lumpur. M.K. was supported by a grant to the Centre for Microbial Diversity and Evolution from the Tula Foundation, the ERD fund 'Centre for Research of Pathogenicity and Virulence of Parasites' (no. CZ.02.1.01/0.0/0.0/16_019/0000759), and Purkyně Fellowship Czech Acad. Sci. N.A.T.I. was supported by an NSERC Canadian Graduate Scholarship.
Chapter 4: Plastid evolution in deep-branching apicomplexans

4.1 Introduction

The Apicomplexa is a phylum of obligate animal parasites including agents of significant human disease such as malaria (Plasmodium spp.) and toxoplasmosis (Toxoplasma gondii), and core symbionts of corals (Seeber et al. 2013; Kwong et al. 2019). They are abundant parasites in nature, with over 6,000 species described and thousands more likely yet to be discovered (Votýpka et al. 2017). Apicomplexan-like parasitism has arisen at least four times in parallel from a free-living plastid-bearing ancestor (Janouškovec et al. 2019; Mathur et al. 2019). In each case, the parasite morphology has converged around the use of an ancestral 'apical complex' structure, which was originally used for feeding but was coopted for infection (Dos Santos Pacheco et al. 2020). Likewise, during each transition to parasitism, the chloroplast underwent convergent reduction, giving rise to a reduced, non-photosynthetic chloroplast, known as the apicoplast. Since its surprising discovery (Wilson et al. 1996; McFadden et al 1996), the apicoplast has been thoroughly investigated as a potential drug target for apicomplexan diseases, and for clues into the evolutionary origins of apicomplexans (Ralph et al. 2001). This reduced organelle is responsible for the essential biosynthesis of isoprenoids, fatty acids and iron-sulphur clusters (Lim and McFadden 2010). While the evolutionary origin of the apicoplast was previously contentious (Keeling 2010), it is now known to be a secondary, red-algal derived plastid that shares a common ancestor with the peridinin-containing plastids found in the sister group to apicomplexans, the dinoflagellates (Janouškovec et al. 2010).

The apicoplast genome is highly reduced, and has become a model for genome evolution in cryptic organelles. Across apicomplexan lineages, the gene content of apicoplasts has proven to be remarkably conserved: they have lost all genes encoding proteins that function directly in photosynthetic electron transfer (i.e., photosynthesis-related genes), they retain genes of non-photosynthetic function, and their genomes are thought to be maintained due to the retention of a small number of non-housekeeping genes (Janouškovec et al. 2010). One major exception are the recently-described corallicolids (coral symbionts) whose plastids contain four genes involved in
chlorophyll biosynthesis in addition to the traditional gene repertoire (Kwong et al. 2019). Phylogenies of plastid-encoded proteins place the corallicolids at the base of the Apicomplexa, which suggests that this may be an ancestral state that was simply lost early in other apicomplexan lineages. In contrast, the nuclear small subunit rRNA gene and mitochondrial phylogenies place the corallicolids closer to the Coccidia, suggesting a more complex history of apicoplast gene loss. A similar incongruence between plastid and nuclear phylogenies was recently observed in the mutualistic apicomplexan, *Nephromyces* (Muñoz-Gómez et al. 2019), and both studies suggested that the currently poor sampling of plastid data from deep-branching and diverse apicomplexan lineages may be a reason for conflicting phylogenetic signals.

To gain further insights into the evolutionary history of the apicoplast, and plastid evolution more generally, we performed whole-genome shotgun (WGS) and transcriptome sequencing surveys on three understudied, deep-branching apicomplexan species, *Aggregata octopiana, Merocystis kathae* and *Margolisiella islandica*. Using phylogenomics, we present a robust multi-gene apicomplexan phylogeny incorporating all published apicomplexan taxa to date (Janouskovec et al. 2019; Mathur et al. 2019; Muñoz-Gómez et al. 2019). We show that species of the genera *Aggregata, Merocystis, Margolisiella*, together with *Rhytidocystis*, are part of a previously unrecognized, monophyletic group of marine invertebrate-infecting apicomplexans that is sister to the Haemosporidia, Piroplasmida, and Coccidia. We also reconstruct the complete apicoplast genomes and plastid metabolism of all three species, in addition to that of four other deep-branching apicomplexan species, *Siedleckia nematoides, Eleutheroschizon duboscqi, Rhytidocystis* sp. 1 and *Rhytidocystis* sp. 2. We find that the apicoplasts of *Aggregata, Merocystis, Siedleckia*, and *Eleutheroschizon* closely resemble other known apicomplexans in gene content and structure. However, the apicoplast genomes of *Margolisiella* and *Rhytidocystis* spp. differ from all known apicoplasts, in that they are more severely reduced, divergent, and have lost the highly-conserved plastid-encoded RNA polymerase (rpoBC) operon.
4.2 Results and Discussion:

A resolved multiprotein phylogeny of the Apicomplexa

We generated new transcriptomes and WGS sequencing data for *Merocystis kathae* and *Margolisiella islandica*, and WGS data from *Aggregata octopiana*, after isolating the parasites from their marine mollusc hosts (SRA PRJNA645464). *Margolisiella islandica* is known to infect Icelandic scallops (*Chlamys islandica*) where it causes an intracellular infection in the heart auricle (Kristmundsson et al. 2011), *Aggregata octopiana* primarily infects the gastrointestinal tract of the common octopus (*Octopus vulgaris*) with various intermediate crustacean hosts (Gestal et al. 1999; Castellanos-Martínez et al. 2013; Castellanos-Martínez et al. 2019), and *Merocystis kathae* infects the renal tissues of the common whelk (*Buccinum undatum*) with intermediate life stages in scallops (Kristmundsson and Freeman 2018) (Figure 11). Host tissue infected with oocysts were dissected and washed to isolate parasite sporocysts and sporozoites from which RNA and DNA were extracted and sequenced for transcriptome and WGS analysis.

![Figure 11: Light micrographs of oocysts of the species sequenced.](image)

Left to right, *Merocystis kathae*, *Margolisiella islandica*, and *Aggregata octopiana*.

To place these species within a phylogenomic context, we added them to a dataset of slow-evolving nuclear genes previously used to resolve deep phylogenetic relationships within the Apicomplexa (Mathur et al. 2019). This dataset was also expanded to incorporate 11 other recently published transcriptomes (Janouskovec et al. 2019; Muñoz-Gómez et al. 2019). The final phylogenetic matrix included 55 taxa, 194 conserved, nucleus-encoded genes, and 58,611
amino acid sites (Appendix Table C1). Maximum-likelihood phylogenies using an empirical profile mixture model (LG+C40+F+G4), and Bayesian analyses using the CAT-GTR model (chain bipartition discrepancies: max diff. = 0.017) (Lartillot et al. 2009; Stamatakis 2014) produced congruent topologies that were well resolved and statistically supported at most internal nodes (Figure 12). The resulting phylogenomic tree confirms the polyphyletic distribution of apicomplexan parasitism, with at least four origins. *Digyalum* is robustly sister to *Platyproteum*, together forming the ‘Squirmida’ (Cavalier-Smith 2014), a group sister to all apicomplexans and chrompodellids (chromerids and colpodellids). *Nephromyces* is sister to the hematozoa (Muñoz-Gómez et al. 2019) and the gregarines (eugregarines and archigregarines) form a fully-supported monophyletic group. The position of *Cryptosporidium* remains problematic: in these analyses it is recovered as sister to the gregarines, but with somewhat weaker support.

**A new apicomplexan class, Marosporida, that infects marine invertebrates**

*Aggregata, Merocystis* and *Margolisiella* all branch with *Rhytidocystis* in a robustly supported monophyletic group (Figure 12). The recovery of *Aggregata* and *Merocystis* as sister taxa is congruent with traditional taxonomic studies and 18S rRNA small subunit gene phylogenies, which have led to their placement in the family, Aggregatidae (Patten 1935; Kristmundsson and Freeman 2018) (Figure C1). However, the Aggregatidae has been placed within the Coccidia, which is not consistent with the phylogenomic tree. Similarly, the placement of *Margolisiella* as the sister taxon to *Rhytidocystis* has also been observed previously in rRNA phylogenies, although with variable statistical support (Kristmundsson et al. 2011; Miroliubova et al., 2020 (Figure C1). Historically, however, *Rhytidocystis* and *Margolisiella* have typically been classified into separate coccidian families, Agamococcidiorida and Eimeriidae, respectively (Levine 1979; Desser and Bower 1997; Leander and Ramey 2006), and very recently proposed to be a new subgroup, Eococcidia, based on concatenated rRNA operon phylogeny, which did not include *Aggregata* or *Merocystis* (Miroliubova et al., 2020). Here, we show with robust multiprotein phylogenomics that these taxa are sisters, but are not coccidians (Figure 12).
The past taxonomic treatments of all these organisms are complex and contradictory. Indeed, the entire apicomplexan lineage is in need of a major revision to better reflect conclusions from molecular and phylogenomic analyses. To best represent their relationships and avoid the confusions of names representing contradictory taxonomic proposals, we propose a new apicomplexan Class, Marosporida, named to reflect the marine nature of the currently recognized members. Within this group, we propose existing subgroups that do not contradict the phylogenetic relationships can be retained: the Aggregatidae is therefore transferred from the Coccidia to the Marosporida to reflect the sister relationship of Aggregata and Merocystis, and similarly the Rhytidocystidae, which was erected for the genus Rhytidocystis (Levine 1979; Leander & Ramey 2006; Rueckert & Leander 2009), can also be transferred to the Marosporida. The Eococcidia (Miroliubova et al., 2020) is also consistent with current phylogenomics, and could be transferred to the Marosporida, although it carries with it the potentially misleading reference to Coccidia. The Agamococcidiorida is an extremely problematic group that will need to be revisited and perhaps abandoned. This group originally contained Rhytidocystidae and Gemmocystidae, which included one member: Gemmocystis cylindrus, a coral-infecting species that was suggested from histology to be closely related to Rhytidocystis (Upton and Peters 1986). Gemmocystis is now often hypothesized to be related to a broader group of coral-infecting apicomplexans, the corallicolids (Kwong et al. 2019). This cannot be tested with molecular data since none was produced for Gemmocystis, but we can conclude that corallicolids are not related to Rhytidocystidae (Kwong et al. 2019; Miroliubova et al., 2020). Whether the corallicolids are also members of Marosporida is still an open question. Corallicolid transcriptomic data remains unavailable, however, 18S rRNA and mitochondrial gene phylogenies do not support this grouping (Appendix Figure C1 and C2; Miroliubova et al., 2020), which, together with the lack of data from other key taxa like Pseudoklossia and Adelina, highlight the need for comparable sequencing data from additional apicomplexan lineages.
Figure 12: Phylogeny of the Apicomplexa and placement of the new class, Marosporida.

A maximum likelihood tree of apicomplexans and their relatives based on 195 nucleus encoded protein markers and 58,611 amino acid sites. Newly sequenced species from this study are shown in bold. Circles at the nodes correspond to non-parametric bootstrap support (RAxML, 1,000 replicates, LG+F+G4 model) and Bayesian posterior probabilities (PP) (PhyloBayes, consensus of two independent runs, GTR+CAT model). All nodes shown have a PP of 1 unless otherwise indicated. The list of proteins used in the phylogenetic matrix, missing data proportions, and the BUSCO completeness of the newly sequenced species can be found in Supplementary Table C1. * denotes that nephromyces is a chimeric taxon assembled from several most closely related lineages from inside the renal sac of a single host.
Taxon sampling does not improve congruence between apicoplast and nuclear phylogenies

Strongly conflicting signals between apicoplast-encoded and nuclear gene phylogenies have been observed in recent publications (Kwong et al. 2019; Muñoz-Gómez et al. 2019), which is unexpected given their shared evolutionary history. One explanation for the incongruence is that deep-branching apicoplast genomes are poorly sampled, making phylogenetic reconstructions less reliable. To fill this gap, we re-analyzed the apicoplast phylogeny with significantly greater taxonomic diversity. To obtain apicoplast genomes from *Aggregata*, *Merocystis* and *Margolisiella*, we conducted whole-genome shotgun (WGS) sequencing using DNA from parasite sporocysts isolated directly from host tissue. In addition, we also assembled a number of complete or near-complete apicoplasts from other previously reported transcriptome data that contained plastid genes, including those of *Rhytidocystis* sp.1, *Rhytidocystis* sp. 2, *Eleutheroschizon*, *Siedleckia*, *Selenidium* and *Digyalum* (Janouškovec et al. 2019). We used hidden Markov models (HMMs) to comprehensively search these transcriptomes for 40 apicoplast-encoded proteins using alignments curated for plastid phylogenomic analyses (Mathur et al. 2019). Apicoplast-encoded protein sequences were filtered and concatenated resulting in a phylogenetic matrix consisting of 58 taxa, 22 proteins and 5,759 amino acid sites (Supplemental Table C2). Using this matrix in combination with ML and Bayesian phylogenetic analyses, we recovered a poorly resolved phylogeny (Figure 13).

The plastid phylogeny, even with the addition of 10 new deep-branching apicomplexans, remains poorly supported and incongruent with the nuclear phylogeny, specifically with respect to the branching order of the major groups (Figure 13). Both phylogenies fully support the sister relationships between *Merocystis* and *Aggregata*, and *Margolisiella* and *Rhytidocystis*. The plastid phylogeny also recovers a monophyletic grouping of the Hematozoa, Piroplasmida, and Coccidia. However, the positions of the gregarines, *Selenidium* and *Siedleckia*, as well as *Nephromyces*, *Hepatozoon*, and *Eleutheroschizon* are not resolved. Interestingly, the position of corallicolids as the sister to all other apicomplexans in the plastid phylogeny, is fully supported in agreement with previous analyses with less diversity (Kwong et al. 2019). The phylogeny was repeated excluding plastid genes extracted from transcriptome data (Appendix Figure C3), which
did not improve the support. We also progressively removed fast evolving sites from the phylogenomic matrix, and tested the stability of the poorly-supported nodes (Supplemental Figure C4). The node placing *Aggregata* and *Merocystis* sister to all apicomplexans other than corallicolids, remains stable, as does support for *Eleutheroschizon* branching basal to the Coccidia (which is also consistent with its position in the nuclear topology). All other deep nodes with poor support have low and fluctuating bootstrap support with the progressive removal of fast-evolving sites indicative of phylogenetic artefacts. Overall, the significant augmentation of apicoplast diversity does little to resolve the incongruence between plastid and nuclear gene trees. Given the fast-evolving nature of the extremely AT%-rich apicoplast genomes, together with the fact that far fewer genes are available in the plastid genome for phylogenomic analyses (5,759 sites in the plastid tree compared to 58,611 sites in the nuclear tree), we conclude, in agreement with the findings of Muñoz-Gómez *et al.*, that the apicoplast-based analyses are less robust in resolving phylogenetic relationships within the Apicomplexa.

**Figure 13: Apicoplast-encoded gene phylogeny.**
A maximum-likelihood tree of apicomplexans based on 22 plastid-encoded genes and 5,759 amino acid sites. Branch support values are inferred from 500 non-parametric bootstraps (IQ-TREE model LG+F+R7) and are indicated by shaded circles on the nodes. Nodes with support less than 70% have been collapsed into polytomies. The shading corresponds to the groupings coloured in Figure 11. * denotes taxa that only have plastid genome data available.
Apicoplast genomes in Margolisiella and Rhytidocystis are highly reduced and lack RNA polymerase genes

The Aggregata and Merocystis apicoplast genomes are extremely similar in gene content, synteny and size (Figure 14A). They contain compact (~38kbp) circular-mapping genomes with an inverted repeat including the 5S, 16S, and 23S rRNAs and the ribosomal protein gene rps4, like apicoplasts of Coccidia and corallicolids. They lack all genes involved in photosynthesis, including the four chlorophyll biosynthesis genes found in the corallicolids (chlL, chlN, chlB and acsF). The only significant differences between the two genomes are the presence of the ribosomal protein gene rpl11 and the RNA polymerase subunit rpoC2A being split in two open reading frames in Merocystis but not Aggregata. Therefore, their apicoplast genomes are overall extremely similar to each other in both structure and gene content, and do not differ substantially from the apicoplast genomes found in the Coccidia and Haemosporidia (Figure 14B).

Unlike Merocystis and Aggregata, the plastid genome of Margolisiella is strikingly reduced compared to all other apicoplasts sequenced to date (Figure 14A and B). The genome is very small (18 kbp), with a strong AT% bias (13.3% GC). The genome is circular, extremely compact, and contains a single copy of the 16S and 23S rRNA genes, and a reduced complement of 18 tRNA genes and 13 ribosomal proteins, along with a single copy of the tufA, clpC and sufB proteins, three ribosomal protein pseudogenes, and two hypothetical proteins (Figure 14A and B). Margolisiella also uses an alternate genetic code where UGA (the “opal” stop codon in the standard genetic code) encodes tryptophan, which is also found in the corallicolids and the two chrompodellids, Piridium and Chromera (Janouškovec et al. 2010; Kwong et al. 2019; Mathur et al. 2019). Unlike all other known apicoplasts, the Margolisiella apicoplast genome has lost all four of its plastid-encoded RNA polymerase genes, which are presumed to be solely responsible for the transcription of the apicoplast genome and therefore are functionally indispensable (Nisbet and McKenzie 2016).
Figure 14: The apicoplast genomes of Aggregata octopiana, Merocystis kathae, and Margolisiella islandica.

(A) Complete apicoplast genomes sequenced in this study. Ribosomal rRNAs are shown in red, ribosomal proteins are shown in brown, and other proteins and tRNAs are shown in blue and pink, respectively. The species name, genome size and genetic content are indicated in grey circles within each genome. Margolisiella lacks plastid-encoded RNA polymerase genes (rpoB, rpoC1, rpoC2). (B) Plastid gene content and structure comparison between parasitic apicomplexans. Presence is denoted by filled circles and gene colours correspond to the gene categories. Dashed circles represent pseudogenes and numbers within the circles represent genes that have been duplicated. * indicates genomes that were assembled from mining of transcriptome sequencing reads. (C) Reconstruction of apicoplast anabolic pathways.

A similarly AT-rich fraction of sequence reads was also observed by Janouškovec et al. in two species of Rhytidocystis, where the authors found apicoplast proteins in their transcriptome data (Janouškovec et al. 2019). Organellar genomes with their high copy number and elevated expression levels can be highly represented not only in genome sequences, but also in transcriptomes if their AT-rich transcripts are enriched by the poly-A selection step (Smith 2013). To determine whether the rhytidocystid apicoplast genomes resembled that of Margolisiella, we mined the publicly available transcriptomes from Rhytidocystis sp. 1 (which infects Travisia forbesii) and Rhytidocystis sp. 2 (which infects Ophelia limacina), for plastid sequences and were able to assemble complete apicoplast genomes. These searches also revealed numerous plastid contigs that allowed for the assembly of complete circular genomes from Siedleckia nematoides and Eleutheroschizon duboscqi (Appendix Figure C5) and fragmented genomic contigs that included most of the expected genes from Selenidium and Digyalum (see Materials and Methods) (Appendix Table C2).

The Rhytidocystis apicoplast genomes are even more reduced than Margolisiella (13 and 14kbp, respectively). They are extraordinarily AT-rich, with a GC content of 11.6% in Rhytidocystis sp. 1, and 9% in Rhytidocystis sp. 2. These are the most AT-rich plastid genomes sequenced to date, and Rhytidocystis sp. 2 even surpasses the AT-richness of the holoparasitic plant, Balanophora.
Although the two species are in the same genus, their apicoplast genomes show considerable divergence. *Rhytidocystis* sp. 1 is more reduced, and encodes only six ribosomal proteins, the 16S and 23S rRNAs, 4 tRNAs, *clpC* and *sufB*, whereas *Rhytidocystis* sp. 2 encodes seven ribosomal proteins (in addition to three that are pseudogenized), the 16S and 23S rRNAs, nine tRNAs, *clpC*, and *sufB* (Figure 14B). Strangely, *Rhytidocystis* sp. 1 uses an alternate genetic code (UGA encodes tryptophan), but *Rhytidocystis* sp. 2 uses the standard genetic code (Figure 14B). Both genomes lack all genes for RNA polymerase, but interestingly have also lost the translation elongation factor, *tufA*, which is present in all other apicoplast genomes sequenced to date. The extreme compaction of the *Rhytidocystis* and *Margolisiella* apicoplast genomes demonstrates that genome reduction has not reached an “end point” in the majority of apicoplasts, despite the appearance of little variation from the best-studied groups, and further emphasizes the likely importance of only two genes, *sufB* and *clpC*, as a barrier to outright loss of the apicoplast genome (Janouškovec et al. 2015).

**The enigmatic transcription of Margolisiella and Rhytidocystis plastid genes**

The lack of plastid-encoded RNA polymerases in *Margolisiella* and *Rhytidocystis* raises the question of how their apicoplast genes are transcribed. We first explored the possibility that the RNA polymerase genes had been transferred to the nucleus and that the plastid-derived polymerase proteins are imported back to the organelle, as many other plastid proteins are, and dinoflagellate plastid RNA polymerases are (Mungpakdee et al. 2014). We used a combination of BLAST and HMMs (Altschul et al. 1990; Finn et al. 2011) to comprehensively search for RNA polymerase proteins based on domain structure and sequence composition in *Margolisiella* transcriptome and WGS data, and *Rhytidocystis* transcriptomes, but found no homologs, despite identifying plastid-encoded polymerase proteins (*rpoB* and *rpoC*) in all other apicoplast-bearing apicomplexans (Figure 15A). By comparison, we also searched for the “missing” *tufA* protein in *Rhytidocystis*, and found *tufA* transcripts from both species with canonical plastid targeting leaders, indicating that *tufA* has been transferred to the nucleus and that its protein product is targeted back to the apicoplast.
Another possible explanation, for which there is a precedent, is that the ancestral plastid-derived RNA polymerase has been lost entirely, and apicoplast transcription relies on a separate and distinct nuclear-encoded polymerase derived from some other source. Land plant plastids, for example, use two different RNA polymerases: a nuclear-encoded polymerase related to homologues from T7 bacteriophage that transcribes non-photosynthesis genes, and the ancestral plastid polymerase that transcribes photosynthesis related genes (Liere et al. 2011). Intriguingly, some holoparasitic plants that have lost many or all photosynthesis related genes have also lost their functional plastid-encoded RNA polymerase, and in the genus *Cuscata* it has been demonstrated that all transcription is now carried out by the phage-derived polymerase (Krause et al. 2003; Krause 2008). To see if such an alternative polymerase exists in *Margolisiella* and *Rhytidocystis*, we searched their transcriptomes as well as the predicted proteins of other plastid-bearing apicomplexans with Pfam HMMs and identified all proteins containing domains associated with the two largest RNA polymerase subunits, phage-type RNA polymerases, and other viral RNA polymerases (El-Gebali et al. 2018) (Figure 15). We identified subunits of the eukaryotic RNA polymerases (RNAPI, II and III) in all the apicomplexans searched, as well as a few phylogenetically ambiguous proteins that were not associated distinctly with a certain polymerase (labelled as ‘uncertain’) (Figure 15A).
Figure 15: The RNA polymerases (RNAPs) present in plastid-bearing apicomplexans, *Digyalum*, and *Piridium*.

(A) The presence of eukaryotic RNAPs I, II and III and other RNAPs with uncertain phylogenetic association, plastid encoded RNA polymerases (PEP) and viral RNA polymerases are represented as portions of the circles. Empty circles indicate that no proteins were found. (B) The two domains of the bacteriophage-derived T7 polymerase (mitochondrial) are represented by circles. A line joining the circles signifies a complete T7 polymerase where the N-terminal domain is attached to the rest of the protein, and no line indicates fragmented proteins. Numbers within the circles denote the number of unique proteins identified.
These searches also found T7 bacteriophage derived RNA polymerases in all taxa, except for *Merocystis* (Figure 15A), which may be due to the incompleteness of the *Merocystis* transcriptome (refer to Appendix Table C1 for BUSCO completeness scores). Most eukaryotic mitochondria use a T7 phage polymerase for transcription of mitochondrial genes, and such polymerases have been found in the genomes of both dinoflagellates and apicomplexans (Li et al. 2001; Teng et al. 2013). All T7 polymerases that we recovered were found to be homologous to these mitochondrially-targeted polymerases. In *Selenidium, Siedleckia,* and *Eleutheroschizon* we retrieved truncated proteins that lack the complete N-terminal domain, while in *Toxoplasma, Piridium* and *Digyalum* we recovered the complete protein (Figure 15B). Intriguingly, we found additional T7 polymerases in several apicomplexans, including *Rhytidocystis.* Two N-terminal domain fragments were found in *Rhytidocystis* sp. 1 while in *Rhytidocystis* sp. 2 we found two truncated T7 polymerase transcripts that are missing the N-terminal domain (Figure 15B). *Margolisiella* contained two non-overlapping T7 polymerase fragments, but it was not clear if they are part of one protein or two different proteins. It is possible that a T7 phage-derived polymerase might be targeted to the apicoplast, or even that the mitochondrial T7 polymerase is dually targeted in these species. A precedent comes from land plants, where a dually targeted RNA polymerase with an ambiguous targeting sequence allows a mitochondrial T7 polymerase to be imported into both the chloroplast and mitochondria (Hedtke et al. 2000). Another possibility is that mitochondrial T7 polymerase may contain a ‘twin’ targeting sequence, represented by a mitochondrial and a chloroplast targeting sequence in tandem. This is seen in the protoporphyrinogen oxidase II enzyme in spinach, which has two in-frame initiation codons, and thus two different proteins are made by alternative translation where the longer protein is targeted to the chloroplasts and the shorter one to the mitochondria (Watanabe et al. 2001). Based on present sequencing data, we cannot convincingly identify the protein responsible for transcription of the *Margolisiella* and *Rhytidocystis* apicoplast-encoded genes, but we hypothesize that an unrecognized but ancient redundancy in apicomplexan RNA polymerases exists and that some single-subunit polymerase, such as a T7 phage polymerase, is targeted to these apicoplasts.
The canonical apicoplast biosynthetic function is conserved in Marosporida

Given the variability in apicoplast genomes in Marosporida, we explored the diversity of organelle function in *Aggregata*, *Merocystis*, and *Margoliisiella*. Generally, apicoplasts are involved in biosynthesis of isoprenoids (MEP), fatty acids (FASII), iron-sulphur (Fe-S) clusters, and part of the tetrapyrrole (heme) biosynthesis pathway (Sheiner et al. 2013). Most apicomplexans retain genes for all four pathways, however, the piroplasms and ‘Symbiont-X ’ only carry out isoprenoid biosynthesis (Janouškovec et al. 2015; Janouškovec et al. 2019), and the marine gregarine clade that includes *Pterospora*, *Lankesteria* and *Lecudina* only retain fatty acid biosynthesis (Mathur et al. 2019).

All plastid homologs for enzymes in these pathways were identified by HMM searches with previously curated alignments (Mathur et al. 2019; see Materials and Methods). We found evidence for the presence of all four pathways in the Marosporida (Figure 14C, Appendix Table S3). In *Margoliisiella*, we found the complete fatty acid biosynthesis pathway, a near complete Fe-S cluster assembly pathway, and homologs for several enzymes involved in both isoprenoid (*ispC, ispE, ispG*) and heme biosynthesis (*hemE* and *hemH*). In *Aggregata*, we found the complete isoprenoid biosynthesis and near complete fatty acid pathway, as well as most enzymes involved in the Fe-S and heme pathways. We also found homologs from all pathways except the MEP pathway in *Merocystis*, however none were complete, probably because the *Merocystis* transcriptomic data was not sequenced as deeply. We then searched for putative targeting leaders indicated by signal and transit peptides at the N-terminus of these plastid-targeted proteins using predictions from SignalP v5.0 and ChloroP v1.1 (Emanuelsson et al. 1999; Dyrløv Bendtsen et al. 2004). We found plastid targeting N-terminal signatures in 3 of these proteins (Appendix Table C3), including components of the Fe-S and FASII pathways. Interestingly, we also found evidence for the cytosolic type I fatty acid synthase in *Margoliisiella* and *Aggregata*. This pathway has been lost in many apicomplexans, but is also retained in the Coccidia, suggesting that both Coccidia and Marosporida retain both type I (cytosolic) and type II (plastidic) fatty acid biosynthesis pathways (Mazumdar and Striepen 2007).
4.3 Chapter 4 conclusions

Here we present a well-resolved phylogenetic framework of the Apicomplexa that includes nearly all recognized apicomplexan groups. This facilitated the identification of a new clade of diverse apicomplexans previously classified into several distinct subgroups, the Marosporida. Whole genome shotgun and transcriptome sequencing shows that plastid metabolisms of this new group is conserved, but the apicoplast genome structure and content are highly variable. Apicoplasts from *Margolisiella* and *Rhytidocystis* have the smallest, most reduced, and most AT-biased genomes known to date, and have distinctively lost plastid-encoded RNA polymerase genes. Taken together with the recent discovery of chlorophyll biosynthesis genes in corallicolids and the intra-genus variation of the *Rhytidocystis* plastids, the patterns of gene retention and loss in deep-branching apicomplexans is more complex than previously thought, as are other relatively stable characters, such as non-canonical genetic codes. This variability will likely continue to increase with greater taxon sampling, given that we only have a handful of representatives from most lineages, and entirely lack complete apicoplast genomes from several key taxa, such as the archigregarines and squirmids. We also still lack nuclear genomic resources for several important groups, such as corallicolids and adeleids. Although our overall understanding of reductive plastid evolution and the emergence of parasitism in the Apicomplexa is still challenged by both gaps in taxon and subcellular compartment sampling, the unexpected genetic diversity and complex evolutionary patterns that have been revealed here and in other recent studies bring us closer to a comprehensive understanding of apicomplexan biology and evolution.
**Taxonomic description**

**Marosporida** Mathur, Kristmundsson, Gestal, Freeman, & Keeling 2020


**Etymology:** ‘Maro ’refers to the marine environment that the hosts of these parasites inhabit.

**Reference phylogeny:** Fig. 1, this paper. *Aggregata octopiana* is closely-related to *Merocystis kathae*, and *Margolisiella islandica* to *Rhytidocystis* sp. 1 and 2.

**Comments:** This clade is inferred from molecular phylogenies. This is a zoological name above level of order and as such falls outside the zoological (and botanical) codes of nomenclature.

### 4.4 Materials and Methods

**Sample collection and DNA/RNA extraction**

*Merocystis kathae* was isolated from a common whelk (*Buccinum undatum*) and *Margolisiella islandica* was isolated from an Iceland scallop (*Chlamys islandica*) that were collected by dredging in Breiðafjörður Bay, Iceland (65° 7.576’N; 22° 44.738’W). Prior to sampling, both the whelks and the scallops were sedated using 0.1% MgSO$_4$ in seawater for 1–2 hours. The renal organ of the whelks were examined under a dissecting microscope for the presence of *Merocystis* infections. Subsequently, the relatively large gamogonic stages of *Merocystis* were retrieved by gently squeezing the renal organ with pointed forceps until the parasites were released. The resulting exudate was collected into concave glass spot plates containing filtered seawater and rinsed with autoclaved seawater three times to remove as much host tissues and mucous as possible. The auricles of the scallops were excised under a dissecting microscope, and infections of *Margolisiella* (all life stages present) examined from a small drop of haemolymph from the auricles. Small samples from heavily infected heart auricles, and its haemolymph, were taken for molecular analysis. DNA and RNA from the samples, infected with *Merocystis* and *Margolisiella*, was extracted using a QIAGEN AllPrep DNA/RNA Mini Kit (Cat. No. 80204).
*Aggregata octopiana* was isolated from a pool of five infected octopuses (*Octopus vulgaris*) collected using traps by local fishermen in Ria de Vigo, Spain (24º 14.09’N, 8º 47.18’W). *Aggregata* oocysts were observed as white spots on the digestive tract with light microscopy. The oocysts were extracted from the caecum and intestine, and light microscopy and histology were used to analyse the morphology and dimensions of the fresh sporocysts.

For DNA extraction, the digestive tract tissues infected with sporogonic stages of the parasite were dissected and homogenized in 10 μL of filtered seawater (FSW) with 1% Tween80 using an electric tissue grinder (IKA-Ultra Turrax T-25). Tissue homogenates were filtered twice with a nylon mesh of 100 μm and 41 μm respectively, to remove tissue fragments. The filtrate was then centrifuged at 1000 g for 5 min. The sporocysts were cleaned using a density gradient centrifugation method according to Gestral et al. (Gestal et al. 1999), and counted in a Neubauer chamber to standardize the sample at 2 x10⁶ sporocyst/μL. Sporocysts were resuspended in 500 μL of extraction buffer (NaCl 100mM, EDTA 25mM pH 8, SDS 0.5%) and disrupted by wet bead milling using a Retsch Mixer Mill MM 300 grinder to release sporozoites. After Proteinase K (Sigma) digestion (1 mg μL⁻¹; 37 ºC overnight), genomic DNA was purified using a phenol:chloroform:isoamil alcohol extraction method (Sambrook et al. 1989). DNA was precipitated with ethanol and sodium acetate overnight at -20 ºC. The precipitated pellet was resuspended in 30 μL of sterile water.

**Genome sequencing, plastid genome assembly, and annotation**

DNA samples from *A.octopiana*, *M.kathae*, and *M.islandica*, were prepared for whole-genome shotgun sequencing at The Centre for Applied Genomics in The Hospital for Sick Children using the Illumina TruSeq Nano kit. The resulting libraries were sequenced on the Illumina HiSeq X sequencer using 150 bp paired-end reads. The quality of the raw reads was assessed using FastQC and trimmed to remove sequencing adaptors with Trimmomatic v0.36 (Andrews; Bolger et al. 2014). Due to a high level of animal host contamination in the samples, the raw reads were first assembled with the metagenomic assembler, MEGAHIT v1.1.4-2 (Li et al. 2015). This initial assembly was used to search for apicoplast contigs using BLAST against known apicoplast genome sequences (Altschul et al. 1990). The raw reads were then mapped onto these
apicoplast contigs of interest and extracted using Bowtie v2.2.6 and BlobTools bamfilter (Laetsch et al.; Langmead et al. 2009). These extracted reads were used for the final apicoplast genome assemblies with Spades v3.11.1 (Bankevich et al. 2012). NOVOPlasty v2.6.3 was used to assemble the inverted repeat regions and close (circularize) the apicoplast genomes (Dierckxsens et al. 2016). The plastid genomes of *Siedleckia nematoides*, *Eleutheroschizon dubosqci*, and *Rhytidocystis* sp.1 and sp.2 were assembled by data mining publicly available RNA-Seq data (SRA PRJNA557242). *Siedleckia* was assembled using NOVOPlasty v2.6.3, and *Eleutheroschizon* and *Rhytidocystis* were assembled based on contig overlaps from assemblies of raw transcript reads using rnaSPAdes 3.13.2 (Bushmanova et al. 2019). The apicoplast genomes were annotated manually in Geneious Prime (www.geneious.com/prime/) and ORFs larger than 100 amino acids were predicted, followed by BLAST searches against the NCBI GenBank non-redundant (nr) databases (Agarwala et al. 2018). rRNA genes were annotated based on predictions made by RNAmer v1.2 using the “Bacteria” setting and tRNAs were annotated using tRNAscan-SE 2.0 (Lagesen et al. 2007; Chan and Lowe 2019).

**Plastid phylogenomic analyses**

A dataset comprising 40 plastid-encoded proteins was compiled based on a previously published dataset (Mathur et al. 2019) and enriched with all apicomplexan plastids that have been sequenced as of January 2020 (Appendix Table C2) as well as the plastid-encoded proteins of *A. octopiana*, *M. kathae* and *M. islandica*. Profile hidden Markov models (HMMs) searches with the above-mentioned protein alignments were also used to identify plastid-encoded proteins from transcriptomic data published by Janouškovec et al., 2019. Hits from the HMM search were aligned with MAFFT v7.222 and poorly aligned regions were removed using trimAl v.1.2 (-gt 0.8) (Capella-Gutierrez et al. 2009; Katoh and Standley 2013). Maximum likelihood trees were made with FastTree v.2.1.7 using the default options (Price et al. 2010). The resulting phylogenies were inspected manually to remove contaminants and paralogs. The selected proteins were then added to the final protein alignments for phylogenomic analyses (see Appendix Table C2 for the list of taxa and proteins). The final protein alignments were aligned with MAFFT v7.222 using the --auto option and trimmed with trimAl v.1.2 (-gt 0.6) (Capella-Gutierrez et al. 2009; Katoh and Standley 2013). These alignments were then filtered so that they
contained only a maximum of 26% missing OTUs and concatenated using SCaFoS v1.2.5 (Roure et al. 2007). The resulting concatenated alignment consists of 22 genes spanning 5759 amino acid positions from 58 taxa. The phylogenomic maximum likelihood analyses were done in IQ-TREE v.1.5.4 with the model LG+F+R7 and 500 non-parametric bootstraps. This model best fits the data according to the Akaike information criterion (AIC) and the corrected Akaike information criterion (AICc) as determined by ModelFinder (Kalyaanamoorthy et al. 2017). Fast evolving site removal was done using site rates generated in IQTREE v.1.5.4 (-wsr option) (Nguyen et al. 2015).

**Transcriptome sequencing and assembly**

Reverse transcription of RNA samples from *M.kathae* and *M.islandica* was carried out using the Smart-Seq2 protocol (Picelli et al. 2014). The cDNA concentration was quantified on a Qubit 2.0 Fluorometer (Thermo Fisher Scientific Inc.). Prior to high-throughput sequencing, 1 µL of the final cDNA product was used as a template for a PCR amplification of the V4 region of the 18S rRNA gene using Phusion High-Fidelity DNA Polymerase (New England Biolabs, Thermo Scientific) and the general eukaryotic primer pair TAREuk454FWD1 and TAREukREV3 (Stoeck et al. 2010). The PCR product was then sequenced by Sanger sequencing. The SSU rRNA gene sequences were used to confirm species specificity and avoid animal host contamination using BLASTn to look for similar sequences in the non-redundant NCBI database (Johnson et al. 2008). Sequencing libraries were then prepared using the Nextera XT protocol, and sequenced on the Illumina MiSeq sequencer using 250 bp paired-end reads. The raw Illumina sequencing reads were merged using PEAR v0.9.6, and FastQC was used to assess the quality of the paired reads (Andrews; Zhang et al. 2014). The adapter and primer sequences were trimmed using Trimmomatic v0.36 and the transcriptomes were assembled with Trinity v2.4.0 (Grabherr et al. 2011; Bolger et al. 2014). The contigs were then filtered for animal host contamination using BlobTools in addition to BLASTn and BLASTx searches against the NCBI nt database and the SWISS-PROT database, respectively (Agarwala et al. 2018; Bateman 2019). Coding sequences were predicted using a combination of TransDecoder v3.0.1 and similarity searches against the SWISS-PROT database (Haas et al. 2013; Bateman 2019). The completeness of the
transcriptomes were assessed with BUSCO v4.0.6 using the alveolate marker gene set (Simão et al. 2015) (Appendix Table C1).

**Nuclear phylogenomics analyses**

Transcriptome data from *M. kathae*, *M. islandica*, and recently published apicomplexan transcriptomes by Janouškové et al. (SRA PRJNA557242) and Muñoz-Gómez et al. (SRR8618777) were added to our dataset. The transcriptomes were searched using BLASTp for a set of 263 genes that have been previously used for apicomplexan phylogenomic analyses and that represent a wide range of eukaryotes (Altschul et al. 1990; Burki et al. 2016; Mathur et al. 2019). The hits were filtered using an e-value threshold of $1 \times 10^{-20}$ and a query coverage of 50%.

In addition to this transcriptomic dataset, the genomic reads of *A. octopiana* and *M. kathae* were also searched for the 263 genes using tBLASTn with an e-value threshold of $1 \times 10^{-20}$. The complete regions of the contigs that contained hits were extracted and coding regions were predicted using Exonerate v.2.2.0 and TransDecoder-v5.1.0 (Slater and Birney 2005; Haas et al. 2013). The final 263 gene alignments were then aligned using MAFFT L-INS-i v7.222 and trimmed using trimAl v1.2 ($-gt 0.8$) (Capella-Gutiérrez et al. 2009; Katoh and Standley 2013). Single gene trees were then constructed to identify paralogs and contaminants using RAxML v8.2.12 (PROT-GAMMA-LG model) with support from 1,000 bootstraps (Stamatakis 2014). The resulting trees were manually viewed in FigTree v1.4.3 and contaminants and paralogous sequences were identified and removed (Rambaut 2007). The final cleaned gene-sets were filtered so that they contained only a maximum of 40% missing OTUs and then concatenated in SCaFoS v1.2.5 (Roure et al. 2007). The resulting concatenated alignment consisted of 194 genes spanning 58,611 amino acid positions from 54 taxa. The phylogenomic maximum likelihood tree was constructed with the heterogenous mixture LG+C40+F+G4 model as implemented in IQ-TREE (Quang et al. 2008; Nguyen et al. 2015). Statistical support was inferred using 1,000 bootstrap replicates using the LG+F+G4 model in RAxML (Stamatakis 2014). The Bayesian tree was computed in PhyloBayes v4.1 using the GTR-CAT model with constant sites removed from the analyses (Lartillot et al. 2009). Four independent chain were run for 10,000 generations and two chains converged with maxdiff = 0.017, while two chains got stuck at local maxima. However, all four chains recovered the same topology in regards to the support of the
Marosporida clade with posterior probability of 1. Furthermore, the chains that recovered the same topology as the best tree had higher log likelihoods.

**RNA polymerase analysis**

To assess the presence and absence of RNA polymerases (RNAPs) in *Margolisiella* and *Rhytidocystis* sp. 1 and sp. 2, we searched genomic and transcriptomic protein predictions from plastid bearing apicomplexans using PFAM HMMs (E < $10^{-5}$, incE < $10^{-5}$, domE < $10^{-5}$) to identify proteins containing domains associated with the two largest RNAP subunits, Rpa1/Rpb1/Rpc1/RpoC (PF00623, PF04983, PF04990, PF04992, PF04997-PF05001), and Rpa2/Rpb2/Rpc2/RpoB (PF00562, PF04560, PF04561, PF04563, PF04565-PF04567, PF10385), as well as phage-type/mitochondrial RNAPs (PF00940, PF10385), and other viral RNAPs (PF00680, PF00978, PF00998, PF02123, PF07925, PF17501) (for a total of 24 PFAM domains) using HMMER v3.1 (Finn et al. 2011; El-Gebali et al. 2018) The same searches were conducted against the SWISS-PROT database to identify non-apicomplexan outgroups (Bateman 2019). Identified SWISS-PROT and apicomplexan proteins were aligned using MAFFT v.7.222 using the PFAM seed alignments as references (Katoh and Standley 2013). The resulting alignments were then trimmed using trimAl v1.2 (-gt 0.3) before being used to generate maximum likelihood phylogenies using FastTree v2.1.3 (Capella-Gutiérrez et al. 2009; Price et al. 2010). To identify which polymerase complexes these proteins corresponded to (e.g., eukaryotic RNAPI, RNAPII, RNAPIII, or prokaryotic RNAP), proteins were annotated using BLASTp searches against the SWISS-PROT database (max_target_seqs 1, E < $10^{-5}$) and their phylogenetic context was interpreted in FigTree. Phylogenetically ambiguous proteins that were not clearly associated with a certain polymerase were labelled as "uncertain".

**Search for plastid-derived biosynthetic proteins**

Profile HMMs were used to identify plastid metabolic proteins in our transcriptomes based on previously curated alignments (Mathur et al. 2019). Profile HMMs were generated using these alignments and HMM searches were conducted on all transcriptomes and genomes using HMMER v3.1 and an e-value threshold of 1e-5(Finn et al. 2011). In addition to this transcriptomic dataset, the WGS reads of *Aggregata*, *Merocystis*, and *Margolisiella* were also
searched for the plastid targeted proteins using tBLASTn with an e-value threshold of 1e-20. The complete regions of the contigs that contained hits were extracted and coding regions were predicted using Exonerate v.2.2.0 and TransDecoder-v5.1.0 (Slater and Birney 2005; Haas et al. 2013). All resulting hits were then extracted and incorporated into the original alignments and realigned using MAFFT v7.222 (--auto option). The alignments were then used to generate phylogenies in FastTree v2.1.3 (Price et al. 2010). The phylogenies were manually scanned in FigTree v1.4.2 and contaminants, paralogs, mitochondrial sequences, and long-branching divergent sequences were identified and removed (Rambaut 2007). The remaining sequences were realigned and used to generate maximum likelihood phylogenies in IQ-TREE v.1.6.9 (Nguyen et al. 2015). Phylogenetic models were selected for each tree individually based on Bayesian Information Criteria using ModelFinder as implemented in IQ-TREE, and statistical support was assessed using 1,000 ultrafast bootstrap pseudoreplicates (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017). SignalP v5.0 and ChloroP v1.1 were used to predict plastid targeting signals on the N-terminal in the proteins recovered (Emanuelsson et al. 1999; Dyrløv Bendtsen et al. 2004). Refer to Appendix Table C3 for all plastid targeted proteins and localization signals recovered.

**Mitochondrial and 18S ribosomal small subunit gene phylogenies**

The three mitochondria-encoded proteins, *cox1*, *cox3* and *cob*, were extracted using BLAST searches against the transcriptomes and WGS assemblies (Altschul et al. 1990). Single proteins alignments were generated using MAFFT v7.222 (--auto option) and trimmed using trimAl v1.2 (-gt 0.6) (Capella-Gutierrez et al. 2009; Katoh and Standley 2013). Proteins were concatenated in Geneious Prime v 2020.1.1. The phylogeny was constructed in IQ-TREE with the LG+F+I+G4 model and 1,000 ultrafast bootstraps. The best-fit model was chosen according to the Bayesian information criterion (BIC) using Model Finder (Kalyaanamoorthy et al. 2017). Nuclear 18S rRNA genes were extracted using BLAST searches against the transcriptomes and genomes (Altschul et al. 1990). Genes were aligned with MAFFT v7.222 (--auto option) and trimmed using trimAl v1.2 (-gt 0.6) (Capella-Gutierrez et al. 2009; Katoh and Standley 2013). Phylogenies were constructed in IQ-TREE with the GTR+G+I model and 1,000 ultrafast bootstraps.
4.5 Chapter-specific acknowledgments

We thank Sunita Sinha (UBC Sequencing Center) and Sergio Pereira (SickKids, Toronto) for technical help regarding Illumina sequencing. We also thank Racquel Singh for help with samples and Martin Kolisko for valuable discussions. This work was supported by grants from the Canadian Institutes of Health Research (MOP-42517) and the Gordon and Betty Moore Foundation (https://doi.org/10.37807/GBMF9201) to PJK. V.M. was supported by a University Graduate Fellowship from the University of British Columbia. W.K.K. was supported by an NSERC (Natural Sciences and Engineering Research Council) Postdoctoral Fellowship and N.A.T.I was supported by an NSERC Canadian Graduate Scholarship.
Chapter 5: Organellar metabolism in the gregarines and squirmids

5.1 Introduction

Apicomplexans are diverse parasites that infect a wide range of animals, and exhibit a large variation in infection strategies, morphology, life cycles, and pathogenicity, across the group (Votýpka et al. 2017). Despite these differences, all apicomplexans have adapted to living permanently within animal hosts, and share the ability to exploit their host for metabolites. In order to do so, they have drastically reshaped their own metabolic capabilities by retaining, losing or gaining genes for metabolic enzymes. Substantial evidence of this stems from genomic and experimental studies on the medically important apicomplexans, *Plasmodium* and *Toxoplasma* (Tymoshenko et al. 2015; Swapna and Parkinson 2017; Blume and Seeber 2018).

While most metabolism takes place in the cytoplasm, several key metabolic pathways are partitioned to the mitochondria and apicoplast in these parasites. The apicoplast carries out type II fatty acid biosynthesis, isoprenoid synthesis, heme biosynthesis, and the formation of iron-sulfur (Fe-S) clusters (Sheiner et al. 2013; McFadden and Yeh 2017), whereas the apicomplexan mitochondria maintains the Fe-S biosynthesis pathway, tricarboxylic acid (TCA) cycle and respiratory chain (Seeber et al. 2008; Seidi et al. 2018).

An exception to this are the *Cryptosporidium* parasites which have lost the apicoplast entirely, and only retain a reduced mitochondria (mitosome) that lacks a genome (Abrahamsen et al. 2004; Xu et al. 2004). Overall, *Cryptosporidium* parasites have an extremely reduced metabolism and scavenge most essential metabolites from their hosts (Sateriale and Striepen 2016). This genus has now been formally re-classified as a gregarine (a large group of deep-branching apicomplexans that typically infect invertebrates), although this relationship is not fully supported by phylogenomics (Carreno et al. 1999; Leander 2008; Mathur et al. 2019). *Cryptosporidium*'s reduced metabolism shows that variation in important metabolic capacities should be expected across the tree of apicomplexans, but most of this diversity has not been examined, and metabolic functions in gregarine lineages remains particularly enigmatic because
genomic data from the gregarines is sparse (partly because they are not in culture). Recent studies using single-cell transcriptomics provide a promising approach to better understand the gregarines and gain insights into the evolution of parasitism in the Apicomplexa as a whole (Janouškovec et al. 2019; Mathur et al. 2019). A key finding from these studies is that certain apicomplexans previously classified as gregarines actually represent an independently evolved apicomplexan-like parasitic group, known as the Squirmida (Cavalier-Smith 2014; Janouškovec et al. 2019; Mathur et al. 2019). A complementary set of gregarines and squirmid transcriptomes have been characterized, and these provide preliminary evidence for metabolic convergence in addition to the morphological similarities between these two groups.

To better understand the extent of convergent evolution between the squirmids and gregarines, characterize the organellar metabolism of these two groups, and to more thoroughly test the deep phylogenetic relationships of the Apicomplexa, we generated high-quality single-cell transcriptomes from the two squirmids species, *Platyproteum vivax* and *Filipodium phascolosomae*, and four gregarine species, *Polyplicarium translucidae*, *Lankesteria metandrocarpae*, *Trichotokara eunicae* and *Selenidium serpulae*. We integrated these new transcriptomes with all recently published apicomplexan sequencing surveys (Janouškovec et al. 2019; Mathur et al. 2019; Mathur et al. 2020) to create a comprehensive and deeply sampled apicomplexan transcriptomic dataset. Using this dataset, we generated a robust phylogenomic tree that resolves the deep nodes of the Apicomplexa, and supports the monophyly of the gregarines with the exclusion of *Cryptosporidium*, and the phylogenetic placement of the Squirmida as sister group to the apicomplexans and chrompodellids. An in-depth comparison of the apicoplast and mitochondrial metabolism shows a strong convergence in the biosynthetic pathways that are partitioned to these organelles in the archigregarines and squirmids. We also find that two lineages of eugregarines have lost all plastid targeted metabolism (suggesting multiple losses of the plastid), and more unexpectedly also have highly reduced their mitochondrial respiratory chains, probably losing the mitochondrial genome as a result.
5.2 Results and Discussion:

We generated single-cell transcriptomes from uncultivated apicomplexan parasites by isolating trophozoites directly from the gut of their animal hosts using previously described approaches (Mathur et al. 2019). We sequenced six transcriptomes from marine invertebrate infecting taxa: *Filipodium phascolosomae, Platyproteum vivax, Polyplicarium translucidae, Lankesteria metandrocarpae, Trichotokara eunicae* and *Selenidium serpulae* (see materials and methods). The phylogenetic placement of these taxa in relation to the Apicomplexa was examined using phylogenomics with a concatenated alignment of 195 nucleus-encoded proteins that have previously been used in both eukaryote-wide and phylum-level phylogenies (Burki et al. 2016; Mathur et al. 2019). Their positions are resolved with strong statistical support in the resulting maximum likelihood (ML) phylogeny consisting of 62 taxa, including 11 dinoflagellate and ciliate outgroup taxa (Figure 16A).

**Phylogenetic placement of Squirmida confirmed with additional taxon sampling**

*Platyproteum vivax* is placed as sister species to *Platyproteum* sp., which branch together with *Filipodium phascolosomae*. This grouping is consistent with 18S rRNA phylogenies (Rueckert and Leander 2009), and together with *Digyalum oweni* form a fully supported monophyletic group, previously hypothesized and named the Squirmida (Cavalier-Smith 2014). This clade falls at the base of the apicomplexans and chrompodellids, and our phylogeny provides further evidence that they represent an independent origin of apicomplexan-like parasites (Janouškovec et al. 2019; Mathur et al. 2019). To test the robustness of this topology, we carried out fast site removal by discarding fastest evolving sites in the concatenated (195 protein) alignment in 10% increments and assessing changes to the bootstrap support (IQTREE, non-parametric bootstraps, \(n = 500\)) (Nguyen et al. 2015) (Figure 16B). This analysis shows that the phylogenetic position of the Squirmida is consistently supported even with the removal of fast sites, and therefore that this phylogenetic placement is not biased by long branch attraction (LBA) artefacts (Figure 16B).
**Figure 16: Multiprotein phylogeny of the Apicomplexa, chrompodellids and Squirmida.**

(A) Maximum likelihood phylogenetic tree of the Apicomplexa and relatives, with dinoflagellate and ciliate outgroups, based on 195 concatenated protein markers (IQ-TREE, LG+C40+F+G4 model). Species newly sequenced in this study are shown in bold orange lettering. Circles and values at nodes correspond to non-parametric bootstrap support (1,000 replicates, IQ-TREE LG+F+R8 model). Black circles indicate 100% support, other support values are indicated on the nodes. Colored circles specify nodes that are tested for the effect of fast-evolving site removal in

(B) The effect of fast-evolving site removal on the bootstrap support of certain nodes (numbered 1-7) in the apicomplexan phylogeny. Line graph shows progressive removal of fastest-evolving sites in increments of 10% and the corresponding bootstrap support values (based on IQ-TREE 500 non-parametric bootstraps, LG+F+R8 model). Colored lines on the graph correspond to the specific nodes that are tested and shown on the phylogeny (above)


---

**Phylogenomics refines taxonomic relationships within the gregarines**

The phylogenomic tree also recovers high support for internal nodes within the gregarines. *Selenidium serpulae* branches as sister species to *Selenidium pygospionis* with full statistical support, and together branch with *Siedleckia nematoides* (Figure 16A). This topology supports the monophyly of the archigregarines, and their previously observed close phylogenetic relatedness to the blastogregarines (Rueckert and Leander 2009; Janouškovec et al. 2019). The robustness of this topology was also tested with fast site removal, and remained consistently fully supported (Figure 16B). This clade also shares morphological synapomorphies, such as the bending and twisting motility in the trophozoite stage, and the ability to feed via myzocytosis, which are not found in any other gregarine lineage (Figure 16C.e) (Leander 2007; Schrével et al. 2016; Simdyanov et al. 2018).
Lankesteria metandrocarpae branches as sister species to Lankesteria abbotii, and together with Lecudina tuzetae and Pterospora schizosoma, with 100% bootstrap support, and while this clade does not share any obvious morphological characteristics, 18S rRNA gene phylogenies also show strong support for this grouping (Leander et al. 2006). Finally, Polyplicarium translucidae branches with Ancora sagittata forming a clade of capitellid polychaete-infecting gregarines, and together are placed sister to Polyrhabdina sp. with full support (Wakeman and Leander 2013; Simdyanov et al. 2017). However, the position of Trichotokara eunicae as sister to all three taxa is poorly supported. The huge (450 µm), tadpole-shaped trophozoites of Trichotokara (Figure 16C.a) are highly distinctive, and do not share morphological characteristics with any other gregarine lineage (Rueckert et al. 2013). Therefore, sequencing data from their close relatives is key to resolving the phylogenetic position of this highly divergent species.

**Gregarines are monophyletic with the exclusion of Cryptosporidium**

In apicomplexan 18S rRNA phylogenies published to date, gregarines are not recovered as a monophyletic group (Simdyanov et al. 2017), and whether they form a single lineage or multiple ones has been debated. The is likely due to highly divergent 18S rRNA genes in many gregarine species leading to phylogenies that are sensitive to taxon sampling and LBA artefacts. Here, using multiprotein phylogenomics, we recover a monophyletic grouping of the gregarines with the exclusion of Cryptosporidium (Figure 16A). Cryptosporidium is placed as basal to the gregarines, Marosporida, Coccidia and Hematozoa, with moderate bootstrap support. However, the support values for this node increase significantly with the removal of fast sites, and is 100% supported with the removal of 50-70% of fast sites. This suggests that Cryptosporidium parasites are likely not monophyletic with the gregarines, and that better taxon sampling and perhaps the discovery of parasites that are closely related to Cryptosporidium is needed to fully resolve their phylogenetic position within the Apicomplexa.

**Squirmid plastid metabolism is highly convergent to apicomplexans**

We next explored the presence of plastids in our newly sequenced taxa by searching for plastid-localized proteins in their transcriptomes. Due to incomplete nature of transcriptome data, only if all components of a pathway are missing do we characterize the entire pathway as being lost.
Filipodium phascolosomae and Platyproteum vivax retain proteins for isoprenoid, heme, and Fe-S cluster biosynthesis, but have lost all components of the plastidial fatty acid type II (FASII) pathway (Figure 17). This pattern is also found in Platyproteum sp. but not in Digyalum oweni which retains all four pathways, indicating that the loss of FASII is clade-specific (Janouškovec et al. 2019; Mathur et al. 2019). Nonetheless, the dependence of the Squirmida on the same four plastid pathways as the apicomplexans (or a subset of them in some lineages), represents an interesting example of metabolic convergence between the groups.

**Multiple losses of plastids in the gregarines**

The archigregarines, Selenidium serpulae and Selenidium pygospionis, as well as Siedleckia nematoides, also retain components of all four canonical apicomplexan plastid-targeted pathways. The complete plastid genomes of Selenidium pygospionis and Siedleckia nematoides were assembled in a recent study (Mathur et al., 2020) suggesting that the archigregarines maintain the typical apicomplexan plastid structure and function that is also found in the Coccidia, Marosporida, and haemosporidians (Figure 17). We also found that Lankesteria metandrocarpae, similar to Lankesteria abbotti, Lecudina, and Pterospora, has lost all plastid biosynthetic pathways except for FASII (Mathur et al., 2019). Interestingly, this clade of marine gregarines (Lankesteria, Lecudina, and Pterospora), are the only eugregarines that contain any plastid metabolism. Trichotokara eunicae and Polyplacium translucidae, in addition to all other eugregarines sequenced to date, do not contain any plastid-targeted proteins (these includes Gregarina where a draft genome of relatively poor coverage is available). Therefore, based on our phylogenomics tree (Figure 16A), this indicates at least two independent losses of plastids within the gregarines, and at least three plastid losses within the Apicomplexa (at least twice in gregarines and once in Cryptosporidium).
Figure 17: Metabolic reconstruction of isoprenoid, fatty acid, heme, and iron-sulphur cluster biosynthesis in the gregarines and squirmids.

Presence and absence of components of the four plastid-targeted pathways and cytosolic alternatives. Blue boxes indicate presence, and blank spaces indicate absence based on genome data. Species newly sequenced in this study are shown in red bold lettering. Predicted intracellular localization of corresponding proteins is shown on the right and summarize known experimental data (Seeber et al. 2005; Fleige et al. 2007; Kořený et al. 2011).


<table>
<thead>
<tr>
<th>ISOPRENOID</th>
<th>Squirmids</th>
<th>Archigregarines</th>
<th>Eugregarines</th>
<th>Predicted Localization</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEP</td>
<td>Tg</td>
<td>Pv</td>
<td>Fp</td>
<td>Do</td>
</tr>
<tr>
<td>HMGCS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMGCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MVD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FASII</td>
<td>ACC1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FabB/F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fabl (ENR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FabG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FabZ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FabH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FabD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FASII</td>
<td>ACC2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FARS</td>
<td>NitU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SuF1A (HesB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SufB (ycf24)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SufC (ycf16)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SufD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SufE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SufI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fd (PetF)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FNR (PetH)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Plastid | Mitochondria | Cytosol | Plastid (encoded in the plastid genome)
When a plastid is acquired by endosymbiosis, the resulting metabolic redundancy can lead to the loss of the endosymbiont pathway, but also potentially the host (cytosolic) version of the pathway, which makes the host dependent on the plastid. Ancestrally, apicomplexans were already dependent on their plastid for isoprenoid biosynthesis (Janouškovec et al. 2015). To examine how the two plastid-lacking gregarine lineages have overcome this metabolic dependency, as well as how the other redundant pathways were resolved, we searched our transcriptomic dataset using HMMs (Hidden Markov Models) for presence of cytosolic alternatives for the isoprenoid mevalonate (MEV/MVA) and fatty acid (FASI) pathways (Figure 17) (Finn et al. 2011) (See materials and methods). We found no evidence for proteins involved in the cytosolic mevalonate pathway in any apicomplexans, chrompodellids or squirmids, including the gregarines, which that have lost the plastidial MEP/DOXP pathway entirely (Figure 17). Dinoflagellates also only contain the plastid MEP/DOXP pathway, irrespective of whether they are photosynthetic or secondarily non-photosynthetic (Slamovits and Keeling 2008; Fernández Robledo et al. 2011; Gornik et al. 2015), suggesting that the MVA pathway was lost in the myzozoan common ancestor. Therefore, similar to other parasites lacking a de novo isoprenoid biosynthesis such as Cryptosporidium and Hematodinium, these gregarines likely have the ability to scavenge isoprenoids from their animal hosts.

We also searched our transcriptomic dataset for the cytosolic FASI pathway, using PFAM HMMs to retrieve all acetyl-CoA carboxylase (ACC) domain containing proteins (El-Gebali et al. 2018) (see materials and methods). Coccidians, such as Toxoplasma and Eimeria, possess both a plastidic (ACC1) and cytosolic (ACC2) fatty acid synthases (Mazumdar and Striepen 2007). Similarly, we found evidence for both proteins in the archigregarines, S. serpulae and S. pygospionis, further emphasizing their convergence of archigregarine plastids with coccidian plastids (Mathur et al., 2020) (Figure 17). Interestingly, we retrieved the cytosolic ACC2 from all eugregarines lacking plastids, and not in the species that contain the plastid FASII (Lecudina, Lankesteria, and Pterospora). The presence of ACC2 has also been shown in the plastid lacking Cryptosporidium, however experimental work suggests that it acts as a fatty acid “elongase” rather than in de novo synthesis of fatty acids (Zhu et al. 2004; Zhu and Guo 2014). Therefore,
we cannot rule out the possibility that gregarines that have lost their plastids are able to scavenge fatty acids directly from their hosts.

**The eugregarines possess a patchy distribution of ETC components**

With the exception of *Cryptosporidium*, apicomplexans are aerobic, possess mitochondria with tubular cristae, and have mostly canonical mitochondrial energy metabolism. However, like dinoflagellates, their mitochondrial genomes are extremely reduced, with a small size (~6-11kbp) encoding only three proteins: cytochrome c oxidase subunit 1 (*cox1*), *cox3*, both in the electron transport chain (ETC) complex IV, and cytochrome b (complex III) (Feagin et al. 1997).

Primarily based on genome sequencing mining of *Toxoplasma* and *Plasmodium*, the apicomplexan mitochondria has been shown to function in Fe-S cluster biosynthesis, heme biosynthesis (partially localized to the apicoplast), the respiratory chain, and the TCA cycle; however, its role in energy generation and functionality has not been formally established (Seeber et al. 2008). The exception to this, is the anaerobic *Cryptosporidium* parasites which have lost the TCA cycle, respiratory chain, and mitochondrial genome, but maintain the Fe-S cluster biosynthesis pathway (Zhu and Guo 2014). To gain insights into the mitochondrial metabolism of the squirmids and gregarines, we searched for the components of these four pathways using KEGG (Kyoto Encyclopedia of Genes and Genomes) database and reciprocal BLASTp searches (Altschul et al. 1990; Kanehisa et al. 2016) (see materials and methods).

The mitochondria of the squirmids has convergently reduced to the same metabolic functions as apicomplexans. *Platyproteum* and *Filipodium* possess all four pathways, and similar to the apicomplexans and *Vitrella*, have lost the canonical ETC complex I and replaced it with an alternative NADH dehydrogenase (NDH2) (Seeber et al. 2008; Flegontov et al. 2015) (Table 1). The archigregarines (*Selenidium*) also retain the canonical apicomplexan mitochondrial metabolism and the alternative complex I (Table 1). In contrast, however, the eugregarines contain components of the Fe-S cluster biosynthesis, the TCA cycle, alternative complex I, but only a surprisingly patchy distribution of the other ETC complexes (Table 1).
<table>
<thead>
<tr>
<th>Pathway</th>
<th>Enzyme name</th>
<th>Abbreviation</th>
<th>Tg</th>
<th>Pv</th>
<th>Fp</th>
<th>Ss</th>
<th>Sp</th>
<th>Ma</th>
<th>Gn</th>
<th>Pr</th>
<th>Lm</th>
<th>Lt</th>
<th>Te</th>
<th>Pt</th>
<th>As</th>
<th>Cc</th>
<th>Cp</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCA cycle</td>
<td>Citrate synthase 1</td>
<td>CS1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Aconitase</td>
<td>ACN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Isocitrate dehydrogenase</td>
<td>ICDH1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>α-ketoglutarate dehydrogenase E1</td>
<td>α-KGDH E1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>α-ketoglutarate dehydrogenase E2</td>
<td>α-KGDH E2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>α-ketoglutarate dehydrogenase E3</td>
<td>α-KGDH E3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Succinyl-CoA synthetase alpha</td>
<td>SCS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Succinyl-CoA synthetase beta</td>
<td>SCS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Succinate dehydrogenase (Fe–S)</td>
<td>SDH</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Fumarate hydratase</td>
<td>FUM</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Malate dehydrogenase</td>
<td>MDH</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Pyruvate carboxylase</td>
<td>PYC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electron transport chain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dihydroorotate dehydrogenase</td>
<td>DHODH</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Malate: quinone oxidoreductase</td>
<td>MQR</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Alternative oxidase</td>
<td>AOX</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Type II NADH: quinone oxidoreductase</td>
<td>NDH2-II</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Succinate dehydrogenase</td>
<td>SDH</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>complex I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cytochrome c reductase</td>
<td>QCR7, UQCRB</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Cytochrome c reductase hinge</td>
<td>QCR6, UQCRH</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Cytochrome c reductase Fe–S</td>
<td>UQCRFS1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Cytochrome c1</td>
<td>CYC1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Cytochrome b</td>
<td>CYTB</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>complex II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cytochrome c oxidase I</td>
<td>COX 1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Cytochrome c oxidase subunit II</td>
<td>COX 2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Cytochrome c oxidase subunit III</td>
<td>COX 3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Cytochrome c oxidase subunit 5b</td>
<td>COX5B</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Cytochrome c oxidase subunit 6b</td>
<td>COX6B</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Hemoglobin</td>
<td>COX10</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Cytochrome c oxidase subunit 11</td>
<td>COX11</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Cytochrome c oxidase subunit 15</td>
<td>COX15</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Cytochrome c oxidase subunit 17</td>
<td>COX17</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>complex IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATP synthase lipid-binding protein (Fo-c subunit)</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>ATP synthase alpha chain</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>ATP synthase beta chain</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>ATP synthase gamma chain</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>ATP synthase delta subunit</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**BUSCO (%)**

99 89 57 87 95 98 94 95 87 78 61 87 88 88 99

● Mitochondrial genome encoded
Table 1: Reconstruction of the gregarine and squirmid mitochondrial metabolism based on KEGG.

Presence and absence of components of the tricarboxylic acid (TCA) cycle, electron transport chain (ETC), and iron-sulphur cluster biosynthesis pathways in the gregarines and squirmids. Species abbreviations are as follows: Tg = Toxoplasma gondii, Pv = Platyproteum vivax, Fp = Filipodium phascolosomae, Ss = Selenidium serpulae, Sp = Selenidium pygospionis, Ma = Monocystis agilis, Gn = Gregarina niphandrodes, Pr = Polyphagodela sp., Lm = Lankesteria metandrocarpae, Lt = Lecudina tuzetae, Te = Trichotokara eunicae, As = Ancora sagittata, Pt = Polyplicarium translucidae Cc = Cephaloidophora communis, Cp = Cryptosporidium parvum. Species newly sequenced in this study are shown in red bold lettering. BUSCO (%) scored are noted below.

Monocystis and Polyphagodela contain the same suite of mitochondrial metabolism as the archigregarines and other apicomplexans, as does Ascogregarina (shown by mining of its draft genome) (Templeton et al. 2010). However, the marine gregarine clade (Lankesteria, Lecudina, and Pterospora) lack all components of complexes III, IV and V (Table 1). A similar pattern of loss is also found in Gregarina, Cephaloidophora, and Polyphagodela. Furthermore, the draft genome sequencing of Gregarina (Toso and Omoto 2007) did not recover a mitochondrial genome (which would be predicted to contain genes encoding proteins in complex III and IV), further pointing towards the loss of these complexes in Gregarina. Lastly, Trichotokara is also missing complex III and IV, however we found several components of the complex V (ATP synthase), and Ancora is missing all components of complexes II, IV and V, yet a fragment of
the genome-encoded cytochrome c oxidase (\textit{cox1}, complex III) was retrieved from the sequencing reads (Table 1).

The loss of ETC complexes in aerobic eukaryotes is rare, however exceptions have been discovered in recent studies, such as the independent losses of complex III in \textit{Chromera velia} and \textit{Amoebophrya ceratii} (Flegontov et al. 2015; John et al. 2019). However, the loss of more than one ETC complex in these eugregarine species suggests that their mitochondria are unlikely to be a major source of energy in these taxa, although it is possible that a certain level of reducing potential may be generated via the alternative type II NADH dehydrogenase. The gregarines that contain a reduced respiratory chain also significantly differ from the anaerobic \textit{Cryptosporidium} because they still maintain the TCA cycle, and we can speculate that together with some energy generation via a simplified ETC, the TCA cycle constitutes as an essential source for the precursors of several key molecules, such as amino acids or heme. Nonetheless, we also cannot rule out that these complexes might be missing due incompleteness of the transcriptomic data (see BUSCO scores in Table 1) and that a critical next step is to generate genome sequencing surveys for these unique eugregarines with modified respiratory chains. Lastly, the widespread loss of complexes III and IV in many eugregarines leads to the prediction that these taxa have aerobic mitochondrial that lack a genome, which has only recently been suggested for the dinoflagellate \textit{Amoebophyra} (John et al. 2018). This is supported by the absence of mitochondrial transcripts in these transcriptomes: transcripts of mitochondrial-encoded genes are a frequent characteristic of myzozoan transcriptomes even with poly-A selection, however in these taxa transcripts for \textit{cox1}, \textit{cox3}, and \textit{cob} are all missing.

\textbf{Reconstruction of the flagellar apparatus in the gregarines and squirmids}

Flagella, and their associated flagellar root apparatus, form a fundamental component of eukaryotic cells (Yubuki et al. 2016). Sequencing of the genomes of free-living relatives of the apicomplexans, \textit{Chromera} and \textit{Vitrella}, showed that most components of the flagella were present in the apicomplexan ancestor, but that losses occurred as apicomplexan lineages differentiated (Woo et al. 2015). To investigate the presence of flagella in the gregarines and squirmids, and to reconstruct the ancestral state of these two groups of parasites that have
convergently evolved similar morphologies, host preferences and extracellular lifecycles, we searched our transcriptomic dataset for 25 known flagella-specific proteins (19 intraflagellar transport (IFT) and 6 basal body proteins) that were previously curated based on the flagellar proteome of *Chlamydomonas* (Silflow and Lefebvre 2001; Pazour et al. 2005) using HMMs (Finn et al. 2011).

We recovered most components of the flagellar apparatus in both *Platyproteum* sp. and *Platyproteum vivax*, as well as in *Digyalum*, however only two flagellar proteins (IFT46 and Tctex-1) were found in *Filipodium*. This may suggest the loss of the IFT machinery in *Filipodium*, which has a highly derived morphology (Figure 16Cd) (Rueckert and Leander 2009), or that these proteins are missing from our analysis due to transcriptome incompleteness (Figure 18). Overall, this pattern suggests that the other squirmids at least have a flagellated stage in their life cycle that has not been observed as yet (the sexual life cycle of these taxa have never been described), and that the squirmid ancestor possessed a functional flagellum that was likely lost in *Filipodium*. We also retrieved a near complete set of IFT and basal body components in the newly described class of Marosporidae (*Rhytidocystis* and *Margolisiella*) (Mathur et al. 2020). This suggests that this clade, similar to the Coccidia, might possess a biflagellated microgamete that has not been observed as yet (Figure 18) (Francia et al. 2016). In contrast, we did not find any evidence of IFT machinery in the gregarines with the exception of *Siedleckia nematoides* that contains IFT70, IFT88 and IFT144. However, all gregarines possess at least one component of the basal body. This is highly similar to the pattern of loss that is found in *Plasmodium* species (Sinden et al. 2010). The ubiquitous presence of basal body proteins, particularly the SFA protein, across the Apicomplexa and in the independently evolved Squirmida, further suggests that these proteins were repurposed or involved in a core function that was not lost in the transition to parasitism (Francia et al. 2012; Woo et al. 2015; Francia et al. 2016).
Figure 18: Flagellar apparatus reconstruction across apicomplexans, chrompodellids and squirmids.

Presence of 19 intraflagellar transport (IFT) and 5 basal body (BB) proteins across 45 species of apicomplexans, chrompodellids, and outgroup taxa. Heatmap shows counts of protein copies found in each species, with the exception of Tctex-1 which shows only presence and absence data. Radial bar charts at nodes estimate the losses along the evolutionary paths and ancestral states of the flagellar apparatus. Inner bar chart (yellow) indicate presence of BB proteins, and outer bar chart (blue) indicate presence of IFT proteins. BUSCO scores (% of complete and fragmented BUSCOs) are shown as bar graphs on the left. Species sequenced in this study are shown in red lettering.
5.3 Chapter 5 conclusions

The organellar metabolism of the eugregarines is highly reduced, even in comparison to other apicomplexan groups. Here we show at least two losses of plastids within the gregarines, and highly reduced mitochondrial respiratory chains in multiple eugregarine lineages. This may point towards the eventual conversion of the mitochondria in these organisms to mitosomes (as has happened in Cryptosporidium), or maybe a stable adaptation to their niche as extracellular gut parasites. Complete genomes or deep-coverage genome sequencing surveys from this group are critical to better understanding the fate of the organelle genomes in these taxa, and how this ties into functional change in the organelle. In contrast, we find that the organellar metabolism of the squirmids is highly convergent with that of other apicomplexan groups, particularly to the coccidians and archigregarines despite their having an independent transition to parasitism. Altogether this reinforces the growing body of data that suggests the perhaps the myzozoa are predisposed to adopt a parasitic lifestyle, and that many such parallel transitions have occurred across the myzozoan phylogeny.

5.4 Materials and Methods

Single-cell isolation and transcriptome sequencing

*Polyplicarium translucidae* was isolated from the gut of the capitellid polychaete, *Notomastus tenuis*, that was collected at low tide at Boundary Bay, Tsawwassen, British Columbia (B.C), Canada. *Filipodium phascolosomae* and *Platyproteum vivax* were isolated from the gut of the peanut worm, *Phascolosoma agassizii*, collected from Ogden Point, Victoria, B.C., Canada at a depth of 7-10m. *Trichotokara eunicae* was collected from the same location from the gut of the polychaete, *Eunice valens*. *Selenidium serpulae* was isolated from the calcareous tubeworm, *Serpula vermicularis*, collected from the rocky pools of Grappler Inlet near the Bamfield Marine Sciences Centre, Vancouver Island, B.C., Canada. *Lankesteria metandrocarpae* was also collected at the same location from the tunicate, *Metandrocarpa taylori*. All specimens were collected in January 2019.
The guts of the host animals were dissected with fine-tipped forceps under a low magnification stereomicroscopes. Hand-drawn glass pipettes were used to collect individual trophozoites under an inverted microscope. Trophozoites were rinsed at least three times in filtered seawater and stored in 2 µL of cell lysis buffer (0.2% Triton X-100 and RNase inhibitor (Invitrogen)). cDNA was synthetized from a single cell, or a pool of 2-3 cells, using the Smart-Seq2 protocol (Picelli et al. 2014). The cDNA concentration was quantified on a Qubit 2.0 Fluorometer (Thermo Fisher Scientific Inc.). 1 µL of the cDNA was used as a template for a PCR amplification of the V4 region of the 18S rRNA gene using Phusion High-Fidelity DNA Polymerase (New England Biolabs, Thermo Scientific). The PCR product were sequenced by Sanger dideoxy sequencing. BLASTn searches of the 18S rRNA gene sequences were used to confirm the identity of the apicomplexan and avoid animal host contamination (Altschul et al. 1990). Once the identity of the apicomplexan was established, sequencing libraries were prepared with the Illumina Nextera XT protocol, and sequenced on the Illumina MiSeq (2x250bp) sequencer.

Transcriptome assembly and phylogenomics analyses
The raw sequencing reads were merged using PEAR v0.9.8 (Zhang et al. 2014). The adaptor and primer sequences were trimmed using cutadapt v2.10 and the transcriptomes were assembled with Trinity v2.8.5 (Martin 2011; Haas et al. 2013). The contigs were filtered for host contaminants using BlobTools, in addition to blastn and blastx searches against the NCBI nt database and the SWISS-PROT database, respectively (Laetsch et al.; Altschul et al. 1990; Bateman 2019). Coding sequences were predicted using a combination of TransDecoder v5.5.0 and similarity searches against the SWISS-PROT database (Grabherr et al. 2011) Assessment of the quality of the assembly and annotation of the transcriptomes was done with BUSCO v4.0.6 using the alveolate marker gene set (Simão et al. 2015).

The transcriptomes were searched using BLASTp for a set of 263 genes that have been previously used for apicomplexan phylogenomic analyses and that represent a wide range of eukaryotes (Altschul et al. 1990; Burki et al. 2016; Mathur et al. 2019). The hits were filtered using an e-value threshold of 1e-20 and a query coverage of 50%. Single gene trees were then constructed to identify paralogs and contaminants using RAxML v8.2.12 (PROT-GAMMA-LG
model) with support from 1,000 bootstraps (Stamatakis 2014). The resulting trees were manually viewed in FigTree v1.4.3 and contaminants and paralogous sequences were identified and removed (Rambaut 2014). The final cleaned gene-sets were filtered so that they contained only a maximum of 40% missing OTUs and then concatenated in SCaFoS v1.2.5 (Roure et al. 2007). The resulting concatenated alignment consisted of 195 genes spanning 55,369 amino acid positions from 64 taxa. The phylogenomic maximum likelihood tree was constructed with the heterogenous mixture LG+C40+F+G4 model as implemented in IQ-TREE v.1.5.4 (Quang et al. 2008; Nguyen et al. 2015). Statistical support was inferred using 1,000 bootstrap replicates using the LG+F+R8 model in IQ-TREE v.1.5.4. Fast evolving site removal was done using site rates generated in IQ-TREE v.1.5.4 (-wsr option) (Nguyen et al. 2015).

**Search for plastid targeted pathways**

HMMs searches were used to identify plastid metabolic proteins in our transcriptomes based on previously curated alignments for isoprenoid, heme, iron-sulphur cluster and fatty acid biosynthesis pathways (Mathur et al. 2019). Profile HMMs were generated using these alignments and searches were conducted on all transcriptomes and genomes using HMMER v3.1 and an e-value threshold of 1e-5 (Finn et al. 2011). The hits were extracted and incorporated into the original alignments and realigned using MAFFT v7.222 (–auto option). The alignments were then used to generate phylogenies in FastTree v2.1.3 (Price et al. 2010). The phylogenies were manually viewed in FigTree v1.4.2 and contaminants, paralogs, mitochondrial sequences, and long-branching divergent sequences were identified and removed (Rambaut 2014). The remaining sequences were realigned and used to generate maximum likelihood phylogenies in IQ-TREE v.1.6.9 (Nguyen et al. 2015). Phylogenetic models were selected for each tree individually based on Bayesian Information Criteria using ModelFinder as implemented in IQ-TREE, and statistical support was assessed using 1,000 ultrafast bootstrap pseudoreplicates (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017). SignalP v5.0 and ChloroP v1.1 were used to predict plastid targeting signals on the N-terminal in the proteins recovered (Emanuelsson et al. 1999; Dyrløv Bendtsen et al. 2004).
Search for cytosolic alternatives of plastid pathways

To consider the retention of cytosolic alternatives to plastid biosynthesis pathways, we searched our transcriptomes using PFAM HMMs (E < $10^{-5}$, incE < $10^{-5}$, domE < $10^{-5}$) to identify proteins containing domains associated with cytosolic isoprenoid biosynthesis pathway (mevalonate) and fatty acid biosynthesis (FAS) using HMMER v3.1 (Finn et al. 2011; El-Gebali et al. 2018). This includes the following enzymes and corresponding PFAM domains: mevalonate kinase (MVK, PF08544 and PF00288), phosphomevalonate kinase (PMVK, PF04275), hydroxymethylglutaryl-coenzyme A reductase (HMGCS, PF01154 and PF08540), mevalonate 5-diphosphate decarboxylase (MVD, PF18376), hydroxymethylglutaryl-coenzyme A reductase (HMGCR, PF00368), acetyl-CoA carboxylase (ACC, PF08326 and PF01039), and biotin carboxylase (PF02785, PF02785). The same searches were conducted against the SWISS-PROT database to identify non-apicomplexan outgroups (Bateman 2019). Identified SWISS-PROT and apicomplexan proteins were aligned using MAFFT v.7.222 (Katoh and Standley 2013). The resulting alignments were then trimmed using trimAl v1.2 (-gt 0.3) before being used to generate maximum likelihood phylogenies using FastTree v2.1.3 (Capella-Gutiérrez et al. 2009; Price et al. 2010). The phylogenies were viewed in FigTree v1.4.2, and presence or absence of the protein was determined (Rambaut 2014).

Reconstruction of mitochondrial metabolism using KEGG (Kyoto Encyclopedia of Genes and Genomes)

To assign annotations to our curated dataset of apicomplexan transcriptomes, we used the web-based KEGG Automatic Annotation Server (KAAS) to carry out bi-directional best hit BLAST searches against the KEGG alveolate-specific gene dataset (Moriya et al. 2007). The resulting files contained KEGG Orthology (KO) assignments and automatically generated KEGG pathways. We then searched these annotations to reconstruct the mitochondrial TCA cycle, electron transport chain (ETC) and iron-sulphur biosynthesis pathways using the lists of apicomplexan mitochondrial proteins compiled by Seeber et al., as a reference (Seeber et al. 2008). Once putative proteins were identified, we carried out reciprocal best hit BLASTp searches against NCBI nr to discard false positives and animal host and bacterial contaminants (Altschul et al. 1990).
**Flagellar apparatus reconstruction**

To investigate the presence and absence of flagellar protein components across the apicomplexan clade, 25 proteins (19 intraflagellar transport (IFT) and 6 basal body proteins) from 45 species were retrieved from a list previously compiled by Woo et al. (Woo et al. 2015). The putative basal body protein TLP1 was excluded from subsequent analyses as its presence in the flagellar apparatus of *C. velia* could not be verified (Portman et al. 2014). These alignments were used as queries for BLASTp searches against the NCBI nr database to identify more apicomplexan representatives as well as non-apicomplexan outgroups. The results of these searches and the original queries were concatenated, clustered using CD-HIT v 4.6 (setting: —c 0.99), and aligned using MAFFT v 7.222 (setting: —linsi) (Altschul et al. 1990; Li and Godzik 2006; Katoh and Standley 2013). These curated alignments were used for downstream analyses.

A predicted proteome dataset was compiled that included our newly sequenced transcriptomes and other publicly available apicomplexan and chromodellid species. The predicted proteome of *Colpodella angusta* was downloaded from the EukProt database, and the predicted proteomes of *Cryptosporidium parvum*, *Cryptosporidium hominis* and *Gregarina niphandrodes* were retrieved from CryptoDB (Heiges et al. 2006; Richter et al. 2020). HMM searches for each flagellar protein were carried out against this dataset (comprising of a total of 31 species) with HMMER v 3.1 (setting: —E 1e-5). The hits were then re-aligned to the query alignments with MAFFT v 7.222 (setting: —auto) and trimmed with trimAl v 1.2 (setting: —gt 0.2) (Capella-Guttiérrez et al., 2009). Maximum likelihood phylogenies were generated using FastTree2 (Price et al. 2010). Paralogous sequences and contaminants were manually removed by viewing the trees in FigTree v1.4.2 (Rambaut 2014). The copy number of each protein per species was totalled with the exception of the basal body protein Tctex-1 where only presence or absence was determined.
Conclusions

The results presented in this thesis show that the deep-branching and poorly studied lineages of apicomplexans provide key insights into the origins and evolution of the group as a whole, and more broadly, in our understanding of the patterns that underlie reductive organelle evolution.

5.1 Chapter summaries and conclusions

In Chapter 2, I investigated the global distribution and inferred ecology of the Apicomplexan-Related Lineages (ARLs) by expansively searching for ARL plastid SSU genes in large-scale high-throughput bacterial amplicon sequencing surveys. These meta-analyses confirmed that all ARLs are coral reef associated (and not generally to all marine environments), but only a subset are actually associated with coral itself. Most unexpectedly, we found Chromera exclusively in coral biogenous sediments, and not within coral tissue, indicating that it is not a coral symbiont, as currently assumed in the field. In contrast, ARL-V was the most diverse, geographically widespread and abundant of all ARL clades and was strictly associated with coral tissue and mucus. ARL-V was found in 19 coral species in reefs, including azooxanthellate corals at depths greater than 500 m. This work was significant in that it showed that ARL-V represents a widespread symbiont of coral, and highlighted the importance of isolating ARL-V and characterizing the nature of this coral symbiosis.

In Chapter 3, I used single-cell transcriptomics to sequence six uncultivated invertebrate parasites representing poorly studied lineages of apicomplexans (a central aim of this thesis). We showed that despite the surprisingly complex origin of apicomplexans from free-living algae, this transition actually occurred at least three times independently. We found that two genera previously classified within the Apicomplexa, *Piridium sociabile* and *Platyproteum* sp. (Squirmida), form separately branching lineages in phylogenomic analyses. This study also presents the first evidence of plastid-targeted metabolism in the gregarines. Overall, these findings suggest a predilection for both the convergent loss of photosynthesis and transition to
parasitism in the apicomplexans and chrompodellids, resulting in multiple lineages of superficially similar animal parasites.

In Chapter 4, I address the objective to sequence plastid genomes from apicomplexan groups lacking any genomic data, by sequencing the apicoplast genomes (and transcriptomes) of three understudied apicomplexans, *Margolisiella islandica*, *Aggregata octopiana*, and *Merocystis kathae*. Phylogenomic analyses showed that these taxa, together with *Rhytidocystis*, form a new clade of apicomplexans that is sister to the Coccidia and Hematozoa (the lineages including most medically significant taxa). Members of this clade retain plastid genomes and the canonical apicomplexan plastidial metabolism. However, the apicoplast genomes of *Margolisiella* and *Rhytidocystis* are the smallest of any apicomplexan plastid sequenced to date. They are also extremely GC-poor genomes and have lost genes for the canonical plastidial RNA polymerase (RNAPs). This new clade of apicomplexans, for which we proposed a new Class, the Marosporida, occupies a key intermediate position in the apicomplexan phylogeny, and adds a new complexity to the models of stepwise reductive evolution of plastid genome structure in these parasites.

In Chapter 5, I aimed to better understand the organellar metabolism of the squirmids and gregarines, and to more thoroughly test the deep phylogenetic relationships of the Apicomplexa. To do so, we generated single-cell transcriptomes from the two squirmids species, *Platyproteum vivax* and *Filipodium phascolosomae*, and four gregarine species, that include both Eugregarines and Archigregarines. Here, I addressed the central goal of this thesis of creating a robust multi-gene apicomplexan phylogeny which includes representatives from all major apicomplexan groups. This phylogeny resolves the deep nodes of the Apicomplexa, and supports the monophyly of the gregarines with the exclusion of *Cryptosporidium*, and the phylogenetic placement of the Squirmida as sister group to the apicomplexans and chrompodellids. We also carried out in-depth analyses of the plastid and mitochondrial metabolism of the two groups. An important finding of this study is that two lineages of Eugregarines have entirely lost plastid targeted metabolism (suggesting multiple losses of the plastid), and more surprisingly also
possess highly reduced electron transport chain (ETC) complexes suggesting that they might have also lost their mitochondrial genomes.

5.2 Future directions

The discoveries stemming from this thesis, such as reduced apicoplast genomes lacking RNAPs in the Marosporida, the retention of all canonical plastid targeted pathways except FASII in members of the Squirmida, and the loss of mitochondrial respiratory complexes in the eugregarines, highlight the diversity and complexity of reductive organelle evolution across this group of organisms. This variability will likely continue to increase with better taxon sampling as we are still missing data from key apicomplexan groups such as the adeleids and haemogregarines. Furthermore, several lineages including the squirmids, lack complete apicoplast and mitochondrial genomes. Therefore, the next step involves tackling these gaps in sampling and sequencing by generating either complete genomes or deep-coverage genome sequencing surveys across the apicomplexan diversity. This will allow for the substantiation of losses that have been proposed based on transcriptomic data, as well as the reconstruction of the complete cellular metabolism, and key characteristics such as the apical complex, across apicomplexans (as well as in apicomplexan-like parasitic lineages). Nonetheless, the unexpected genetic diversity and complex evolutionary patterns that have been presented in this thesis, and in other recent studies, brings us closer to a comprehensive understanding of apicomplexan biology and evolution.
References


mortalities in numerous commercial scallop populations in the northern hemisphere
OPEN. 8: 7865.

*Margolisiella islandica* sp. nov. (Apicomplexa: Eimeridae) infecting Iceland scallop 


115. Kuvardina ON, Leander BS, Aleshin V V., Myl’nikov AP, Keeling PJ, Simdyanov TG. 
(2002). The phylogeny of Colpodellids (Alveolata) using small subunit rRNA gene 
sequences suggests they are the free-living sister group to apicomplexans. *J Eukaryot 
Microbiol*.


RNAmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 
35: 3100–3108.


sequence clade of Apicomplexans: Description of the marine parasite *Rhytidocystis 
291.

123. Leander BS. (2008). Marine gregarines: evolutionary prelude to the apicomplexan 


patterns of high alpha and low beta diversity in tropical parasitic and free-living protists. 


Appendices

Appendix A - Supplementary Information for Chapter 2

Figure A1. Map distribution of samples analysed in the present study.
Figure A2. Reference phylogeny based on the plastid 16S SSU rRNA gene.
Figure A3. Outlier removal for distribution analysis.
<table>
<thead>
<tr>
<th>Title</th>
<th>Environment</th>
<th>Paper/MEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cold water coral microflora: Lophelia pertusa</td>
<td>Coral</td>
<td>Kellogg et al. 2017</td>
</tr>
<tr>
<td>Patterns of Bacterial Host Associations Suggest Different Ecological Strategies Between Two Reef Building Cold-Water Coral Species</td>
<td>Coral</td>
<td>Muscatine et al. 2016</td>
</tr>
<tr>
<td>16S Amplicons of Microbial Communities of Cold Water coral mound Rockall megaglenops</td>
<td>Coral</td>
<td>van der Bijl et al. 2015</td>
</tr>
<tr>
<td>Coral-Associated Microbes Targeted Leci Environmental</td>
<td>Coral</td>
<td>PRJNA275249</td>
</tr>
<tr>
<td>Microbiome Variation in Corals with Distinct Depth Distribution Ranges Across a Shallow–Mesophotic Gradient (15–65 m)</td>
<td>Coral</td>
<td>Gislén et al. 2017</td>
</tr>
<tr>
<td>The Coral Immune Response Facilitates Protection Against Microbes During Tissue Regeneration</td>
<td>Coral</td>
<td>van der Walt et al. 2018</td>
</tr>
<tr>
<td>16S Amplicon Survey of Caribbean Coral Microbiota</td>
<td>Coral</td>
<td>PRJNA35944</td>
</tr>
<tr>
<td>16S rRNA pyrosequencing of the Mediterranean gorgonian Paramuricea clavata</td>
<td>Coral</td>
<td>Vassie et al. 2013</td>
</tr>
<tr>
<td>16S rRNA Gene Top Pyrosequencing of Corals-Associated Bacteria in Red Sea</td>
<td>Coral</td>
<td>Lee et al. 2012</td>
</tr>
<tr>
<td>Bacterial communities of two colony morphs of the Hawaiian reef coral, Montipora capitata</td>
<td>Coral</td>
<td>Skene-Biglow et al. 2015</td>
</tr>
<tr>
<td>Bacterial species associated with Acropora millepora Raw sequence reads</td>
<td>Coral</td>
<td>Wright et al. 2017</td>
</tr>
<tr>
<td>Comparative Microbial Analysis to Understand Taxonomic Fingerprints of Healthy Coral Holobiont Communities at southern Great Barrier Reef, Australia</td>
<td>Coral</td>
<td>Sjöpelt, J., &amp; Linne, M. C. 2013</td>
</tr>
<tr>
<td>Coral Reef Bacterioplankton Targeted Leco (Leci)</td>
<td>Coral</td>
<td>PRJNA169890</td>
</tr>
<tr>
<td>Depth Driven Patterns and Inter-Domain Microbe-Microbe Interactions within the Microbiome of the Coral Solenastrea sedentaria</td>
<td>Coral</td>
<td>PrJNA556144</td>
</tr>
<tr>
<td>Fringing reef Microbes</td>
<td>Coral</td>
<td>Smbolovsky et al. 2017</td>
</tr>
<tr>
<td>Natural volcanic CO2 inputs reveal future trajectories for heterotrophic associations in corals and sponges</td>
<td>Coral</td>
<td>Marques et al. 2015</td>
</tr>
<tr>
<td>Plesiogorgia carinata holobiont Targeted Leco (Leci)</td>
<td>Coral</td>
<td>Ng et al. 2013</td>
</tr>
<tr>
<td>Coralized bacterial 16S rRNA from Acropora millepora adults</td>
<td>Coral</td>
<td>Ludin, A. K. 2014</td>
</tr>
<tr>
<td>A high-resolution microbial inventory of the Montreux coral reef ecosystem</td>
<td>Coral</td>
<td>McClain et al. 2012</td>
</tr>
<tr>
<td>Characterization of filamentous and bacterial communities associated with in situ, ocean water, and corals tissue samples</td>
<td>Coral</td>
<td>Staley et al. 2017</td>
</tr>
<tr>
<td>Highly Heterogeneous Bacterial Communities Associated with the South China Sea Reef Corals-Patent ultra, Gokcebas learning and Acropora millepora</td>
<td>Coral</td>
<td>Lue et al. 2013</td>
</tr>
<tr>
<td>Negatively Curved Settlement of Corals</td>
<td>Coral</td>
<td>PRJEB16281</td>
</tr>
<tr>
<td>Multidimensional plots of the stony coral Porites asteroidea on pyrosequencing abundance and community composition</td>
<td>Coral</td>
<td>McNally et al. 2017</td>
</tr>
<tr>
<td>The microbiome of coralline macroalgae has a key role in mediating holobiont health and survival upon disturbance</td>
<td>Coral</td>
<td>Gislén et al. 2016</td>
</tr>
<tr>
<td>Future reef declawling under a business-as-usual CO2 emission scenario</td>
<td>Coral</td>
<td>Doré et al. 2013</td>
</tr>
<tr>
<td>Bacterial Communities of Three Mexican Metastunes Lakes in Central Asia</td>
<td>Freshwater</td>
<td>Boehr et al. 2016</td>
</tr>
<tr>
<td>Holobiont dynamics and association networks among like bacterioplankton taxa</td>
<td>Freshwater</td>
<td>Elter et al. 2012</td>
</tr>
<tr>
<td>Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities</td>
<td>Freshwater</td>
<td>Anderson et al. 2010</td>
</tr>
<tr>
<td>Microbial diversity in the deep sea and the underexplored “true biosphere”</td>
<td>Hydrothermal Vent</td>
<td>Sorgo et al. 2006</td>
</tr>
<tr>
<td>Metatransome isolated from a hydrothermal vent</td>
<td>Hydrothermal Vent</td>
<td>Brandeis et al. 2010</td>
</tr>
<tr>
<td>The seasonal structure of microbial communities in the Western English Channel</td>
<td>Seawater</td>
<td>Gibert et al. 2011</td>
</tr>
<tr>
<td>Ecology of the rare microbial biosphere of the Arctic Ocean</td>
<td>Seawater</td>
<td>Galand et al. 2009</td>
</tr>
<tr>
<td>Seawater Targeted Leco (Leci)</td>
<td>Seawater</td>
<td>Compbell et al. 2010</td>
</tr>
<tr>
<td>The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes</td>
<td>Seawater</td>
<td>Kitchman et al. 2010</td>
</tr>
<tr>
<td>Water mass-specificity of bacterial communities in the North Atlantic revealed by massively parallel sequencing</td>
<td>Seawater</td>
<td>Apipata et al. 2011</td>
</tr>
<tr>
<td>A Dissolved Oxygen Threshold for Shifts in Bacterial Community Structure in a Seasonally Hypoxic Estuary</td>
<td>Seawater</td>
<td>Spiteri et al. 2015</td>
</tr>
<tr>
<td>Bacterioplankton Composition and Dynamics in the North Pacific Subtropical Gyre</td>
<td>Seawater</td>
<td>SIA00870</td>
</tr>
<tr>
<td>A Global Census of Marine Microbes (COHMM)</td>
<td>Seawater and Sediment</td>
<td>Andrews et al. 2010</td>
</tr>
<tr>
<td>Bacterial diversity and community composition from seawater to subsurface</td>
<td>Seawater and Sediment</td>
<td>Walsh et al. 2016</td>
</tr>
<tr>
<td>Global Patterns of Bacterial Skin Diversity in Seawater and Seawater Ecosystems</td>
<td>Seawater and Sediment</td>
<td>Jürgen et al. 2011</td>
</tr>
<tr>
<td>Community Structure of Benthic Bacteria In Uplwading Ecosystem</td>
<td>Sediment</td>
<td>SIA00989</td>
</tr>
<tr>
<td>Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean Ridge</td>
<td>Sediment</td>
<td>Jeongman et al. 2012</td>
</tr>
<tr>
<td>Diversity and abundance of estuarine-reducing microorganisms in a Mediterranean lagoon complex (Antivador de Golf, Innareo)</td>
<td>Sediment</td>
<td>Pavlovic et al. 2017</td>
</tr>
<tr>
<td>Shallow Water Marine Sediment Bacterial Community Shifts Along a Natural CO2 Gradient in the Mediterranean Sea Off Valvet, Italy</td>
<td>Sediment</td>
<td>Kartlhi et al. 2014</td>
</tr>
<tr>
<td>Sinking Alklen Influences Benthic Microorganisms</td>
<td>Sediment</td>
<td>Ruff et al. 2014</td>
</tr>
<tr>
<td>Bacterial Community Composition of South China Sea Sediments through Pyrosequencing-Based Analysis of 16S rRNA Genes</td>
<td>Sediment</td>
<td>Zhu et al. 2014</td>
</tr>
<tr>
<td>pH dominates variation in tropical soil archaeal diversity and community structure</td>
<td>Soil</td>
<td>Tripoli et al. 2015</td>
</tr>
</tbody>
</table>
Appendix B - Supplementary Information for Chapter 3

Figure B1. Myxozoan 18S SSU rRNA phylogenetic tree.
Figure B2. Plastid phylogenomic tree.
Table B1. Plastid-targeted proteins and localization signal analysis.

<table>
<thead>
<tr>
<th>Matysprotein</th>
<th>Peridium</th>
<th>Lecudina</th>
<th>Platyproteum</th>
<th>KEY</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>N</td>
<td>Y (incomplete)</td>
<td>Y (incomplete)</td>
<td>N = no protein found</td>
</tr>
<tr>
<td>fabI</td>
<td>N</td>
<td>Y SP + TP</td>
<td>Y (incomplete)</td>
<td>y</td>
</tr>
<tr>
<td>fabG</td>
<td>N</td>
<td>Y SP + TP</td>
<td>Y (incomplete)</td>
<td>N</td>
</tr>
<tr>
<td>fabZ</td>
<td>N</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
</tr>
<tr>
<td>fabI</td>
<td>N</td>
<td>Y SP + TP</td>
<td>Y (incomplete)</td>
<td>N</td>
</tr>
<tr>
<td>fabG</td>
<td>N</td>
<td>Y SP + TP</td>
<td>Y (incomplete)</td>
<td>N</td>
</tr>
<tr>
<td>fabH</td>
<td>N</td>
<td>Y SP + TP</td>
<td>Y (incomplete)</td>
<td>N</td>
</tr>
<tr>
<td>fabD</td>
<td>N</td>
<td>Y SP + TP</td>
<td>Y (incomplete)</td>
<td>N</td>
</tr>
<tr>
<td>fabD</td>
<td>N</td>
<td>Y SP + TP</td>
<td>Y (incomplete)</td>
<td>N</td>
</tr>
<tr>
<td>Isoprenoid Synthesis (MEP/DOXP Pathway)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dxs</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>dxr/ispC</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>ispD</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>ispE</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>ispF</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>ispG</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Isoprenoid Synthesis (MEP/DOXP Pathway)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dxs</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>dxr/ispC</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>ispD</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>ispE</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>ispF</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Heme Synthesis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALAS</td>
<td>Y (incomplete)</td>
<td>Y mitochondrial</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>hemH</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>hemB</td>
<td>N</td>
<td>Y (incomplete)</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>hemF</td>
<td>Y (incomplete)</td>
<td>Y (incomplete)</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>hemC</td>
<td>Y (incomplete)</td>
<td>Y (no prediction complete)</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>hemD</td>
<td>N</td>
<td>Y (incomplete)</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>hemF</td>
<td>Y SP but no TP</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fe-S Cluster Assembly</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISCA-like</td>
<td>N</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Fd</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>sufC</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>sufD</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>sufE</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>sufF</td>
<td>Y (incomplete)</td>
<td>Y SP + TP</td>
<td>N</td>
<td></td>
</tr>
</tbody>
</table>

Table B2. Presence of Metabolic Pathways based on KEGG Modules.
<table>
<thead>
<tr>
<th>KEGG ID</th>
<th>Description</th>
<th>Functional Category</th>
<th>KEGG ID</th>
<th>Description</th>
<th>Functional Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>M00386</td>
<td>Ubiquitin system</td>
<td>Protein metabolism</td>
<td>M00384</td>
<td>Terpenoid backbone biosynthesis</td>
<td>Other Metabolism</td>
</tr>
<tr>
<td>M00383</td>
<td>Spliceosome</td>
<td>RNA processing</td>
<td>M00354</td>
<td>RNA processing</td>
<td>DNA, RNA, Protein metabolism</td>
</tr>
<tr>
<td>M00425</td>
<td>RNA polymerase</td>
<td>DNA, RNA, Protein metabolism</td>
<td>M00397</td>
<td>Repair system</td>
<td>DNA, RNA, Protein metabolism</td>
</tr>
<tr>
<td>M00296</td>
<td>Polyamine biosynthesis</td>
<td>Other Metabolism</td>
<td>M00412</td>
<td>Photosynthesis</td>
<td>Carbohydrate Metabolism</td>
</tr>
<tr>
<td>M00134</td>
<td>Purine metabolism</td>
<td>Amino acid metabolism</td>
<td>M00632</td>
<td>Purine metabolism</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>M00027</td>
<td>Histidine metabolism</td>
<td>Amino acid metabolism</td>
<td>M00531</td>
<td>Glycan biosynthesis</td>
<td>Carbohydrate Metabolism</td>
</tr>
<tr>
<td>M00068</td>
<td>DNA polymerase</td>
<td>Other Metabolism</td>
<td>M00742</td>
<td>Cysteine and methionine metabolism</td>
<td>Amino acid metabolism</td>
</tr>
<tr>
<td>M00368</td>
<td>Central carbohydrate metabolism</td>
<td>Carbohydrate Metabolism</td>
<td>M00002</td>
<td>Cofactor and Vitamin metabolism</td>
<td>Cofactor and Vitamin Metabolism</td>
</tr>
<tr>
<td>M00692</td>
<td>Cell signaling</td>
<td>Other Metabolism</td>
<td>M00170</td>
<td>Central carbohydrate metabolism</td>
<td>Carbohydrate Metabolism</td>
</tr>
<tr>
<td>M00335</td>
<td>Carbon fixation</td>
<td>Other Metabolism</td>
<td>M00142</td>
<td>Central carbohydrate metabolism</td>
<td>Carbohydrate Metabolism</td>
</tr>
<tr>
<td>M00044</td>
<td>Branched-chain amino acid metabolism</td>
<td>Amino acid metabolism</td>
<td>M00023</td>
<td>Central carbohydrate metabolism</td>
<td>Carbohydrate Metabolism</td>
</tr>
<tr>
<td>M00028</td>
<td>Arginine and proline metabolism</td>
<td>Amino acid metabolism</td>
<td>M00360</td>
<td>Energy generation</td>
<td>Other Metabolism</td>
</tr>
<tr>
<td>Monocystis</td>
<td>Other Metabolism</td>
<td></td>
<td>Monocystis</td>
<td>Other Metabolism</td>
<td></td>
</tr>
</tbody>
</table>
Appendix C - Supplementary Information for Chapter 4

Figure C1. Maximum likelihood (ML) 18S rRNA gene phylogeny of the Apicomplexa with dinoflagellates as the outgroup.

18S rRNA ML phylogeny
Model = GTR+I+G

- 100% bootstrap
- ≥ 94% bootstrap

KF018659 Plasmodium ovale
XR_002273101 Plasmodium falciparum
AF013418 Theileria parva
U27498 Alexandrium margaelefi
AF280076 Voronomas pontica
JN986792 Chromera velia
HM245049 Vitrella brassicaformis
AF022194 Gymnodinium fuscum
FJ587219 Karenia brevis
AF288023 Hematodinium sp

0.1

AF176837 Hepatozoon catesbiana
MC4522531 Hepatozoon griseisciuri
JN11811571 Hepatozoon spedon
KP8813491 Hemolivia stellata
MH6150061 Hepatozoon canis
HQ2249591 Haemogregarina bali
HQ2249611 Hemolivia mariae
HQ2249581 Dactylosomes ranarum
AF494059 Adelina bambarooniae
JN2276681 Margoliisella islandica

Margoliisella islandica (TRINITY_DN24884_c2_g1)

AF176837 Hepatozoon catesbiana
MC4522531 Hepatozoon griseisciuri
JN11811571 Hepatozoon spedon
KP8813491 Hemolivia stellata
MH6150061 Hepatozoon canis
HQ2249591 Haemogregarina bali
HQ2249611 Hemolivia mariae
HQ2249581 Dactylosomes ranarum
AF494059 Adelina bambarooniae
JN2276681 Margoliisella islandica

Margoliisella islandica (TRINITY_DN24884_c2_g1)

AF176837 Hepatozoon catesbiana
MC4522531 Hepatozoon griseisciuri
JN11811571 Hepatozoon spedon
KP8813491 Hemolivia stellata
MH6150061 Hepatozoon canis
HQ2249591 Haemogregarina bali
HQ2249611 Hemolivia mariae
HQ2249581 Dactylosomes ranarum
AF494059 Adelina bambarooniae
JN2276681 Margoliisella islandica

Margoliisella islandica (TRINITY_DN24884_c2_g1)

AF176837 Hepatozoon catesbiana
MC4522531 Hepatozoon griseisciuri
JN11811571 Hepatozoon spedon
KP8813491 Hemolivia stellata
MH6150061 Hepatozoon canis
HQ2249591 Haemogregarina bali
HQ2249611 Hemolivia mariae
HQ2249581 Dactylosomes ranarum
AF494059 Adelina bambarooniae
JN2276681 Margoliisella islandica

Margoliisella islandica (TRINITY_DN24884_c2_g1)
Figure C2. Phylogenetic placement of *Aggregata*, *Merocystis*, and *Margolisiella* based on mitochondria-encoded proteins.
Figure C3. Phylogenetic placement of *Aggregata*, *Merocystis*, and *Margolisiella* based on 22 plastid-encoded proteins.
Figure C4. The effect of fast-evolving site removal on bootstrap support in the plastid phylogeny.
Figure C5. Complete circular apicoplast genomes.
Table C1. Nuclear-encoded phylogenomic statistics.

<table>
<thead>
<tr>
<th>Number of genes</th>
<th>BUSCO lineage database: <em>alveolata</em>odb10</th>
<th>Margasiella islandica</th>
<th>Mercystis kathae</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes present (%)</td>
<td>Complete BUSCOs (C)</td>
<td>Complete single-copy BUSCOs (S)</td>
<td>Complete and duplicated BUSCOs (D)</td>
</tr>
<tr>
<td>Aggregata octopiana</td>
<td>81</td>
<td>65</td>
<td>16</td>
</tr>
<tr>
<td>Alexandrium tamarense</td>
<td>96</td>
<td>45</td>
<td>33</td>
</tr>
<tr>
<td>Alphanema stax</td>
<td>77</td>
<td>61</td>
<td>122</td>
</tr>
<tr>
<td>Ancora spallata</td>
<td>93</td>
<td>171</td>
<td>171</td>
</tr>
<tr>
<td>Ascorhodora alcaica</td>
<td>17</td>
<td>64.3</td>
<td>28.7</td>
</tr>
<tr>
<td>Babesia bovis</td>
<td>58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Babesia micro</td>
<td>76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chlamydomonas communis</td>
<td>85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chroma vella</td>
<td>97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coelodris angusta</td>
<td>29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cryptosporidium hominis</td>
<td>62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cryptosporidium murli</td>
<td>99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cryptosporidium parvum</td>
<td>97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dicytadon platy</td>
<td>98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elmeria tamborina</td>
<td>85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elmeria tenella</td>
<td>78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eltherococcus duboecog</td>
<td>71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Giardia lamblia</td>
<td>88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glielminia foliaceum</td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hymenaea spium</td>
<td>55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Helicospira caprellae</td>
<td>33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Helicoplumum sp</td>
<td>74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Karamia brevis</td>
<td>99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kryptorhodinum foliaceum</td>
<td>88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lecanoderma ruuzi</td>
<td>91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mesoasella asadi</td>
<td>76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mercystis kathae</td>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monocystis agta</td>
<td>94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neospora caninum</td>
<td>95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neoplasma occidentalis</td>
<td>97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oxythys marina</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perulius marinus</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protamococcus asper</td>
<td>59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasmodium berghei</td>
<td>71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasmodium chabaudi</td>
<td>96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasmodium falciparum</td>
<td>95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasmodium knowlesi</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasmodium vivax</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasmodium yoell</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platyphlobom sp</td>
<td>88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plophobom sp</td>
<td>84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasmodium schizansoma</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhytibacysites sp</td>
<td>78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhytibacysites sp2</td>
<td>91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sarcocystis neurona</td>
<td>85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selenidium pygopionia</td>
<td>91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scolices nematodes</td>
<td>89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symbiodinium sp</td>
<td>76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symbiont X</td>
<td>93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theleba annulata</td>
<td>86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Toxoplasma gondii</td>
<td>95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vibrio brassicatum</td>
<td>76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voronomas pontica</td>
<td>21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table C2. Plastid phylogenomics statistics.

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Source (NCBI accession)</th>
<th>% genes present</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregata octopiana</td>
<td>this study</td>
<td>86</td>
</tr>
<tr>
<td>Babesia bovis</td>
<td>NC_011385</td>
<td>88</td>
</tr>
<tr>
<td>Babesia orientalis</td>
<td>NC_028020</td>
<td>91</td>
</tr>
<tr>
<td>Bas nicotiana</td>
<td>NW_021703914</td>
<td>95</td>
</tr>
<tr>
<td>Chromera vella</td>
<td>HM22967</td>
<td>91</td>
</tr>
<tr>
<td>Coreilladoid</td>
<td>MB304845</td>
<td>95</td>
</tr>
<tr>
<td>Cyclospora cayetanensis</td>
<td>KP868208</td>
<td>100</td>
</tr>
<tr>
<td>Cystospora suis</td>
<td>PRJNA341953</td>
<td>91</td>
</tr>
<tr>
<td>Dicyasium ovelli</td>
<td>PRJNA557242</td>
<td>64</td>
</tr>
<tr>
<td>Enteria tenella</td>
<td>NC_004823</td>
<td>95</td>
</tr>
<tr>
<td>Eleutherocystis dubosqui</td>
<td>PRJNA557242</td>
<td>73</td>
</tr>
<tr>
<td>Hammondia hammondi</td>
<td>KL559887</td>
<td>86</td>
</tr>
<tr>
<td>Hepatozoon carinii</td>
<td>MH970886</td>
<td>91</td>
</tr>
<tr>
<td>Leucoctyeson caulleryi</td>
<td>NC_022967</td>
<td>95</td>
</tr>
<tr>
<td>Margoliella islandica</td>
<td>this study</td>
<td>59</td>
</tr>
<tr>
<td>Merocystis kathae</td>
<td>this study</td>
<td>95</td>
</tr>
<tr>
<td>Neospora caninum</td>
<td>MK770338</td>
<td>100</td>
</tr>
<tr>
<td>Nephromyces occidentalis</td>
<td>NC_0573196</td>
<td>95</td>
</tr>
<tr>
<td>Pireodium socialbe</td>
<td>MK962129</td>
<td>95</td>
</tr>
<tr>
<td>Plasmodium berghei</td>
<td>K0233130</td>
<td>100</td>
</tr>
<tr>
<td>Plasmodium chabaudi</td>
<td>AB649423</td>
<td>100</td>
</tr>
<tr>
<td>Plasmodium coatneyi</td>
<td>AB649420</td>
<td>95</td>
</tr>
<tr>
<td>Plasmodium gallinaceum</td>
<td>NC_031963</td>
<td>100</td>
</tr>
<tr>
<td>Plasmodium malariase</td>
<td>NC_031401</td>
<td>91</td>
</tr>
<tr>
<td>Plasmodium ovale</td>
<td>AB649417</td>
<td>95</td>
</tr>
<tr>
<td>Plasmodium falciparum</td>
<td>JG437297</td>
<td>100</td>
</tr>
<tr>
<td>Plasmodium yoelli</td>
<td>NW_018223983</td>
<td>95</td>
</tr>
<tr>
<td>Rhodocyclus sp</td>
<td>PRJNA557242</td>
<td>41</td>
</tr>
<tr>
<td>Sarcocystis neurona</td>
<td>PRJNA279361</td>
<td>91</td>
</tr>
<tr>
<td>Spermatodon pygopliensis</td>
<td>PRJNA557242</td>
<td>55</td>
</tr>
<tr>
<td>Spermatodon nematodes</td>
<td>PRJNA557242</td>
<td>86</td>
</tr>
<tr>
<td>Thelidia equil</td>
<td>NW_004675960</td>
<td>36</td>
</tr>
<tr>
<td>Thelidia parva</td>
<td>NC_027759</td>
<td>66</td>
</tr>
<tr>
<td>Toxoplasma gondii</td>
<td>U87145</td>
<td>95</td>
</tr>
<tr>
<td>Vitrella brassicarum</td>
<td>NC_014345</td>
<td>100</td>
</tr>
<tr>
<td>Pathway</td>
<td>Taxa</td>
<td>Library type</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>---------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Iron-sulfur cluster assembly protein SuD</td>
<td>Aggregata</td>
<td>WGS</td>
</tr>
<tr>
<td>Sulfl family cysteine desulfurase</td>
<td>Aggregata</td>
<td>WGS</td>
</tr>
<tr>
<td>Iron-sulfur cluster assembly protein SuE</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>Iron-sulfur cluster assembly protein SU</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>Iron-sulfur cluster assembly protein SU</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>4-hydroxy-3-methylbut-2-en-1-yl diphosphate MDP/IDP/XP</td>
<td>Aggregata</td>
<td>WGS</td>
</tr>
<tr>
<td>4-hydroxy-3-methylbut-2-en-1-yl diphosphate MDP/IDP/XP</td>
<td>Aggregata</td>
<td>WGS</td>
</tr>
<tr>
<td>4-hydroxy-3-methylbut-2-en-1-yl diphosphate MDP/IDP/XP</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>2-C-methyl-D-erythritol 2,4-cyclophosphate MDP/IDP/XP</td>
<td>Aggregata</td>
<td>WGS</td>
</tr>
<tr>
<td>2-C-methyl-D-erythritol 2,4-cyclophosphate MDP/IDP/XP</td>
<td>Aggregata</td>
<td>WGS</td>
</tr>
<tr>
<td>1-deoxy-D-xylulose-5-phosphate reductoisomerase MDP/IDP/XP</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>Proteorhodopsin oxidase</td>
<td>Heme</td>
<td>WGS</td>
</tr>
<tr>
<td>Ferrochelatase</td>
<td>Heme</td>
<td>WGS</td>
</tr>
<tr>
<td>Coproporphyrinogen III oxidase</td>
<td>Heme</td>
<td>WGS</td>
</tr>
<tr>
<td>Uroporphyrinogen decarboxylase</td>
<td>Heme</td>
<td>WGS</td>
</tr>
<tr>
<td>5-aminolevulinic acid dehydratase</td>
<td>Heme</td>
<td>WGS</td>
</tr>
<tr>
<td>D-3-hydroxycrotonyl-acyl carrier protein dehydrogenase FASII</td>
<td>Aggregata</td>
<td>WGS</td>
</tr>
<tr>
<td>D-3-hydroxycrotonyl-acyl carrier protein dehydrogenase FASII</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>Enoyl-acyl carrier protein reductase FASII</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>Beta-ketocarboxy-acyl carrier protein synthase II FASII</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>Beta-ketocarboxy-acyl carrier protein synthase II FASII</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>Beta-ketocarboxy-acyl carrier protein reductase FASII</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>Beta-ketocarboxy-acyl carrier protein reductase FASII</td>
<td>Aggregata</td>
<td>WGS</td>
</tr>
<tr>
<td>Malonyl-CoA-acyl carrier protein transacylase FASII</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>Malonyl-CoA-acyl carrier protein transacylase FASII</td>
<td>Aggregata</td>
<td>WGS</td>
</tr>
<tr>
<td>Beta-ketocarboxy-acyl carrier protein synthase II FASII</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>Beta-ketocarboxy-acyl carrier protein synthase II FASII</td>
<td>Aggregata</td>
<td>WGS</td>
</tr>
<tr>
<td>1-deoxy-D-xylulose-5-phosphate synthase MDP/IDP/XP</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>5-aminolevulinic acid synthase</td>
<td>Heme</td>
<td>WGS</td>
</tr>
<tr>
<td>Acetyl-CoA carboxylase</td>
<td>Margaritella</td>
<td>WGS</td>
</tr>
<tr>
<td>Acetyl-CoA carboxylase</td>
<td>Aggregata</td>
<td>WGS</td>
</tr>
<tr>
<td>Acetyl-CoA carboxylase</td>
<td>Margaritella</td>
<td>Transcriptome</td>
</tr>
</tbody>
</table>