Reconstruction of Plastid Proteomes of Apicomplexans and Close Relatives Reveals the Major Evolutionary Outcomes of Cryptic Plastids

Varsha Mathur, Eric D. Salomaki, Kevin C. Wakeman, Ina Na, Waldan K. Kwong, Martin Kolisko, and Patrick J. Keeling

1Department of Botany, University of British Columbia, 3156-6270 University Blvd., Vancouver V6T 1Z4, BC, Canada
2Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic
3Institute for the Advancement of Higher Education, Hokkaido University, Sapporo 060-0810, Hokkaido, Japan
4Present address: Department of Biology, University of Oxford, 11a Mansfield Rd., Oxford OX1 3SZ, United Kingdom
5Present address: Instituto Gulbenkian de Ciência (IGC) Rua da Quinta Grande, 6, 2780-156 Oeiras, Portugal

*Corresponding author: E-mail: varsha.mathur@biology.ox.ac.uk.
Associate editor: Crystal Hepp

Abstract

Apicomplexans and related lineages comprise many obligate symbionts of animals; some of which cause notorious diseases such as malaria. They evolved from photosynthetic ancestors and transitioned into a symbiotic lifestyle several times, giving rise to species with diverse non-photosynthetic plastids. Here, we sought to reconstruct the evolution of the cryptic plastids in the apicomplexans, chrompodellids, and squirmids (ACS clade) by generating five new single-cell transcriptomes from understudied gregarine lineages, constructing a robust phylogenomic tree incorporating all ACS clade sequencing datasets available, and using these to examine in detail, the evolutionary distribution of all 162 proteins recently shown to be in the apicoplast by spatial proteomics in Toxoplasma. This expanded homology-based reconstruction of plastid proteins found in the ACS clade confirms earlier work showing convergence in the overall metabolic pathways retained once photosynthesis is lost, but also reveals differences in the degrees of plastid reduction in specific lineages. We show that the loss of the plastid genome is common and unexpectedly find many lineage- and species-specific plastid proteins, suggesting the presence of evolutionary innovations and neofunctionalizations that may confer new functional and metabolic capabilities that are yet to be discovered in these enigmatic organelles.

Key words: apicomplexans, organelle evolution, plastids, parasites, reductive evolution.

Introduction

Apicomplexans are a large group of obligate symbionts of animals known to cause harmful diseases that impact global health and economies, most notably the malaria parasite, Plasmodium (Votýpka et al. 2017). These parasites are ubiquitous in nature, and although the vertebrate parasites attract much more attention, apicomplexans infect an enormous diversity of animals in both terrestrial and marine environments, for example, corals (Kwong et al. 2019), tunicates (Saffo et al. 2010), and a wide variety of other invertebrates (Rueckert et al. 2019). They are highly adapted to surviving within animal cells and body compartments and have reduced metabolic capabilities, relying instead on their host for essential nutrients and metabolites (Piro et al. 2021). Therefore, it was a surprising discovery that most apicomplexans harbor a non-photosynthetic plastid, known as the apicoplast (McFadden et al. 1996). The apicoplast is derived from the secondary endosymbiosis of a red alga (Janouškovec et al. 2010) and no longer performs metabolic functions associated with photosynthesis. However, it still carries out essential non-photosynthetic functions, such as cofactor biosynthesis (e.g., Fe–S cluster and heme), isoprenoid biosynthesis (MEP/DOXP), and fatty acid biosynthesis (FASII) (McFadden and Yeh 2017). These metabolic pathways largely depend on plastid-targeted proteins that are encoded in the nucleus and trafficked to the organelle via a bipartite leader sequence, following synthesis in the cytosol (Waller et al. 2000). At least one metabolically important plastid protein is still encoded in the plastid genome in most species (such as the sufB gene involved in Fe–S cluster assembly), and this reduced plastid genome otherwise mainly encodes information processing genes (Mathur, Kwong, et al. 2021). However, there are exceptions to this such as Cryptosporidium spp., where the plastid has been completely lost (Zhu et al. 2000).

The apicoplast is an important drug target for diseases caused by apicomplexans (McFadden and Roos 1999), but it also raises many questions about the evolutionary history of the group, and how it evolved from free-living...
photosynthetic ancestors into obligate animal parasites. Recent studies aimed at sequencing understudied apicomplexan lineages, in particular performing single-cell transcriptome sequencing of invertebrate-infecting apicomplexans, revealed that this transition to parasitism has occurred many times independently, giving rise to multiple clades of “apicomplexan-like” parasites (Janouškovec et al. 2019; Mathur et al. 2019). In each independent transition to parasitism, photosynthesis has been lost, yet the resulting retention and/or loss of plastid metabolobic pathways, the composition of genes on the plastid genome, and even the retention of the organelle itself appear to be highly variable (Salomaki and Kolisko 2019).

Our present understanding of plastid function in apicomplexans mainly comes from the medically important and experimentally tractable apicomplexans, *Toxoplasma* and *Plasmodium* (Sheiner et al. 2013). Observations from these parasites shape our understanding of plastid reductive evolution not only in other apicomplexans, but even in other eukaryotic lineages with secondarily non-photosynthetic plastids, such as chrysophytes (Dorrell et al. 2019), red algae (Salomaki et al. 2015), and parasitic plants (Krause 2008).

However, based on recent transcriptome sequencing data from diverse apicomplexans and other related parasitic lineages, it is not clear to what extent apicomplexan plastids converged on the same functions or on how much functional diversity has been retained. What becomes of the plastid-targeted proteins in the different apicomplexan lineages? Why are plastid genomes retained versus lost after the loss of photosynthesis? And taking into account new sequencing data from diverse apicomplexan-like parasitic lineages, do the same principles of plastid reduction hold true across the clade at large?

Answering these questions has been difficult given the limited tools available to study the vast majority of apicomplexan diversity, which has meant that only the most obvious and functionally important pathways have been analyzed to date. To address this problem more comprehensively, we take advantage of spatial proteomics data from the hyperplexed Localization of Organelle Proteins by Isotope Tagging (hyperLOPIT) study in *Toxoplasma gondii* (Barylyuk et al. 2020). This study involved separating mechanically disrupted cells on density gradients and identifying proteins belonging to shared subcellular compartments based on the distinct abundance distribution profiles formed by different organelles and cell compartments (Mulvey et al. 2017). This approach gives a more functional and assumption-free catalogue of proteins related to a given compartment, and using this approach, 129 proteins out of a total of 3,832 proteins were assigned to the apicoplast above a 99% probability threshold, with another 33 further assigned to the apicoplast below the high-confidence cutoff. Over 60% of these proteins are newly attributed to the plastid therefore significantly expanding our understanding of the functional capabilities of the organelle (Barylyuk et al. 2020).

Here, we investigate the patterns of plastid reduction in secondarily non-photosynthetic plastids by expansively searching for orthologs of the hyperLOPIT apicoplast data set across 10 lineages and 60 species of apicomplexans, chrompodellids, and squirmids (ACS) (here termed as the “ACS” clade). We also sequenced five new single-cell transcriptomes from the poorly sampled gregarine lineages, *Selenidium*, *Polyplacophoraceae*, and *Cephaloidphora*, to create a more robust dataset for this analysis. We present a 203-gene phylogeny incorporating all publicly available apicomplexan sequencing datasets that provides an evolutionary framework to interpret patterns of photosynthesis loss, plastid reduction, and plastid loss. Finally, we characterize the main outcomes of plastid reduction in this group and discuss the implications of the presence of uncharacterized novel and lineage-specific proteins in these highly reduced organelles.

**Results and Discussion**

**Phylogenomics of Apicomplexans and Related Groups Reveals Multiple Events of Plastid Reduction**

To interpret the patterns of plastid evolution across the group, we first sought to generate a robust phylogeny of the apicomplexan and related alveolate lineages. Recently, there has been a massive increase in new transcriptome sequencing surveys from diverse members of the ACS clade (Janouškovec et al. 2019; Mathur et al. 2019; Mathur, Wakeman, et al. 2021; Salomaki et al. 2021; Yazaki et al. 2021). Combinations of all publicly available ACS datasets at the time that analyses were completed, with extensive alveolate and stramenopile outgroup taxa, resulted in a concatenated alignment of 203 genes and 92 species (fig. 1). The genome of *Porospora gigantea* (Boisard et al. 2022) has since been released and will provide valuable data for future investigation. Overall, both our Bayesian (CAT-GTR + G model) and Maximum Likelihood (ML, LG + C60 + F + G model) phylogenies fully support (100% BS, 1.0 PP) the monophyly of the ACS group, and the monophyly of the following subgroups: chrompodellids, squirmids, hematozoans, coccidians, maresporidians, gregarines, and *Cryptosporidium* (fig. 1).

The phylogenetic position of *Cryptosporidium* is elusive due to conflicting topologies in recent apicomplexan multi-gene phylogenetic analyses (Janouškovec et al. 2019; Mathur et al. 2019; Mathur, Kwong, et al. 2021; Salomaki et al. 2021). Most of these phylogenies do not show strong support for *Cryptosporidium* branching as a sister group to the gregarines, and whether *Cryptosporidium* or the gregarines fall at the base of the Apicomplexa also remains uncertain. Our ML phylogeny provides moderate support (87% BS) for the gregarines branching as the earliest diverging lineage (fig. 1); however, Bayesian inference of our dataset produces a topology where *Cryptosporidium* branches first also with moderate support (0.84 PP) (supplementary fig. S1, Supplementary Material online). Extensive topology testing carried out by Salomaki et al. (2021) suggests that the support for
Fig. 1. Phylogenomic tree of ACS. ML phylogeny of apicomplexans comprised 203 genes and 54,661 sites. PMSF bootstrap support (BS) values (n = 1,000) are shown on the branches. Gregarines species that are newly sequenced in this study are shown in bold. A summary of the distribution of plastids, photosynthesis, and plastid genomes in the ACS clade derived from current literature (Janouškovec et al. 2019, Mathur et al. 2019, 2021, Yazaki et al. 2021) and this study are presented.
Cryptosporidium emerging at the base of apicomplexans is more robust and that the alternate topology is driven by systematic errors. The resolution of earliest diverging apicomplexan lineage is likely impossible without more robust evolutionary models or obtaining data from yet-undiscovered relatives of Cryptosporidium that would break its long branch.

**Phylogenetic Relationships of Gregarines**

Gregarines are a unique group of apicomplexans that almost exclusively infect invertebrates and different lineages possess independently reduced plastid types. We added single-cell transcriptomes from five species representing poorly sampled lineages: Selenidium terebellae, Selenidium orientale, Selenidium pisinnus, Polyplicarium lacrimae, and Cephaloidophora sp. (fig. 1). The genus Selenidium is extremely diverse so these three Selenidium species sequenced are more distantly related to one another than are several other sampled apicomplexans that are treated as different genera or classes (Schrével et al. 2016; Rueckert et al. 2019).

Polyplicarium lacrimae and Cephaloidophora sp. fall within the expected subclades with full support (100% BS, 1.0 PP). Interestingly, Cephaloidophora sp. branches as a sister group to Heliospora caprellae rather than Cephaloidophora communis. This could be an artifact of poor taxon sampling in a very long-branching clade of gregarines. However, recent 18S SSU rRNA phylogenies with better taxon sampling do not suggest that Heliospora or Cephaloidophora are paraphyletic groups (Wakeman et al. 2021). We expect that a better taxon sampling of this group in phylogenomic analyses will be needed to resolve the fine-scale relationships of this divergent clade. Overall, our phylogenomic tree provides support for the monophyly of gregarines, and the relationships within subgroups also remain well-supported and reflect the branching patterns observed in previous studies (Janouškovec et al. 2019; Mathur et al. 2019; Mathur, Wakeman et al. 2021; Salomaki et al. 2021).

One noteworthy exception is within the Selenidium clade, where support values are only moderate. Siedleckia is classified as a blastogregarine and has been placed as a sister group to Selenidium (Janouškovec et al. 2019); however, with better taxon sampling of Selenidium species, random gene resampling, and heteroatousce site removal, we show that the bootstrap support for Siedleckia branching within Selenidium increases, which suggests a potential impact of systematic error influencing the observed topology (supplementary fig. S2, Supplementary Material online). Furthermore, Selenidium and blastogregarines share many ancestral apicomplexan traits that are rare or missing in other gregarines such as feeding via myzocytosis and the retention of plastids (Simdyanov et al. 2018). It is possible that blastogregarines evolved from within the paraphyletic Selenidium; however, blastogregarines are currently represented by only one species. Improved taxon sampling of both groups is likely to help clarify the evolutionary relationships between Selenidium and blastogregarines.

**Evolutionary Conservation of Plastid Proteins Across Apicomplexans and Relatives**

Based on this phylogeny, we confirm that photosynthesis has been lost numerous times independently in the ACS clade giving rise to many cryptic (non-photosynthetic) plastids within the group. Furthermore, our analyses demonstrate that the plastid retained in the ancestor of apicomplexans has been differentially reduced and lost in different apicomplexan lineages. To examine the extent to which these plastids have convergently reduced, we performed Hidden Markov Model (HMM) profile searches (Eddy 2011) for the expanded set of 162 apicoplast proteome proteins identified by hyperLOPIT in Toxoplasma across our 58 species ACS dataset (refer to supplementary table S1, Supplementary Material online for the complete list of taxa analyzed). Most sequencing data available for this group are based on RNA-seq and might represent an incomplete sampling of expressed genes. Therefore, species with BUSCO scores <40% were excluded from final analyses to minimize the effects of this sampling bias (Simão et al. 2015; refer to supplementary table S1, Supplementary Material online for BUSCO scores). Individual ML phylogenies incorporating all hits retrieved were built and manually inspected to remove animal host and bacterial contaminants. Each protein was also functionally annotated using the eggNOG, Pfam, and KEGG databases, to assign the protein to a broad function or specific metabolic pathway (Kanehisa et al. 2016; Huerta-Cepas et al. 2019; Mistry et al. 2021). However, some proteins could not be assigned to any known protein family or functional domain and are annotated as “hypothetical” proteins (supplementary fig. S3, Supplementary Material online).

The presence and absence distributions of the 115 functionally annotated apicoplast proteins in the ACS clade are shown in figure 2, with proteins clustered according to their metabolic functions, and a further 47 uncharacterized apicoplast proteins are presented in supplementary figure S3, Supplementary Material online. Not surprisingly the Toxoplasma apicoplast assigned proteins are most highly conserved in other coccidians as close evolutionary relatives are expected to share more genes than more distant relatives of Toxoplasma; Neospora shares 93% and the protococcidian, Eleutherocschizon shares 64%. However, even within this group, there is variability: 5% of Toxoplasma apicoplast proteins were retrieved only in other coccidia, 7% were restricted to the sarcocystidae clade (Toxoplasma and Neospora), and a further 2% were specific to Toxoplasma, revealing the presence of lineage and species-specific proteins even in highly reduced cryptic plastids.

Other apicomplexan lineages that share the greatest number of Toxoplasma apicoplast proteins are the marosporid (e.g., Rhodocystis [67%]) and hematozoans (e.g.,...
Fig. 2. Presence and absence of putative homologs of 115 *Toxoplasma* apicoplast proteins in the ACS clade. ToxoDB protein numbers (right) and associated protein names are provided in supplementary table S2, Supplementary Material online. The species tree (top) shows phylogenetic relationships and major clades. Proteins are colored and clustered by metabolic function. BUSCO scores as estimates of completeness are shown as a heatmap bar underneath the species names. Proteins that were assigned to the *Toxoplasma* apicoplast with a lower probability threshold by hyperLOPIT are indicated with a *.
Plasmodium [62%]). Interestingly, we also retrieved most of the Toxoplasma apicoplast proteome in deep branching ACS lineages such as the Selenidium (71%) and photosynthetic chromerids (Vitrella 76% and Chromera 62%), suggesting that most of these proteins were present in the shared ancestral plastid of the group. This finding allows us to infer that different degrees of reduction has occurred in various ACS lineages once photosynthesis was lost; for example, the piroplasms, Babesia (51%) and Theileria (47%), retain fewer apicoplast proteins and therefore have undergone further loss unique to those lineages. The lecudinids (Lecudina and Lankesteria), urosporids (Pterospora), and Platyproteum have reduced even further, and both contain only 14% of the Toxoplasma apicoplast proteins. In all other eugregarines and Cryptosporidium, which have lost the apicoplast, we retrieved a maximum of four apicoplast proteins that are involved in either redox homeostasis or protein folding, as well as one protein of unknown function in Cryptosporidium. Barylyuk et al. (2020) showed that the abundance-distribution profiles of the endoplasmic reticulum (ER) proteins and the apicoplast proteins in their hyperLOPIT study were very similar, and given that most apicoplast proteins are trafficked via the ER, this association might reflect a role of these proteins in folding and redox processes during sorting of proteins to the apicoplast (Bidau et al. 2018). Therefore, we predict that the four proteins found in taxa without apicoplasts previously functioned in control of protein trafficking to the apicoplast but are now retained for various other functions in these organisms.

Patterns of Metabolic Functions and Genome Loss in ACS Plastids

Our analyses also reveal the shared general metabolic processes carried out by plastids across the ACS clade (supplementary table S2, Supplementary Material online). Overall, the biosynthesis of isoprenoids, fatty acids (FASII), heme, and Fe–S clusters synthesis are widely present, in addition to a large suite of proteins with functions associated with the plastid genome such as DNA and RNA maintenance, transcription, and translation. Other proteins that are broadly present are involved in carbohydrate, lipid, and phospholipid metabolism, redox homeostasis, protein processing, and folding, in addition to membrane metabolite and protein transporters. We find both variation and independent losses of these pathways in different lineages, for example, Platyproteum, and the piroplasms have lost FASII and carbohydrate synthesis, while the lecudinids and urosporids have exclusively kept these two pathways. This suggests that no one specific pathway drives the retention of plastids, but rather that different parasites might be able to scavenge certain metabolites depending on their host or life cycle (Piro et al. 2021), allowing for the evolutionary loss of different pathways in different lineages.

Another observation is the interdependency of metabolic pathways leading to patterns of shared presence or loss. For example, FASII and carbohydrate metabolism are either both kept or lost across all plastids. Carbohydrate metabolism includes the components of the pyruvate dehydrogenase complex, which converts pyruvate into acetyl-CoA and is essential for the initiation of FASII in Plasmodium (Shears et al. 2015). Similarly, the ferredoxin redox system (Fd/FNR) and isoprenoid biosynthesis are also co-associated and a recent in vivo study in Toxoplasma showed that Fd acts as a critical electron donor to the terminal enzymes of the MEP pathway (Henkel et al. 2022). Most organisms in the ACS clade are neither in culture nor experimentally tractable, so these kinds of patterns are important to suggest specific pathway co-dependencies are widespread and play an important role in shaping the mosaic of metabolic pathways that are retained across reduced plastids.

In some lineages, we did not recover any proteins associated with the plastid genome (a total of 38 proteins involved in DNA and RNA metabolism, transcription, and translation), which strongly suggests that they have lost their plastid genome. This conspicuous pattern allows us to predict plastid genome losses, even based on incomplete transcriptomic data, in Platyproteum, Lankesteria, Lecudina, Pterospora, Colpodella, and “Symbiont X.” There are many hypotheses explaining why nonphotosynthetic plastids retain genomes despite the mass endosymbiotic gene transfer (EGT) of plastid genes to the nucleus, and specifically why only two nonhousekeeping genes, clpC and sufB, remain on most apicoplast genomes (Barbrook et al. 2006; Allen 2017). The “limited transfer window” hypothesis states that EGT depends on organelle lysis and is likely to be lethal to a cell with a single organelle. This suggests that sufB and clpC have simply not had the chance to be transferred to the nucleus and that their retention is a historical accident (Janouškovec et al. 2015; Muñoz-Gómez and Slamovits 2018). Additionally, lecudinids, urosporids, colpodellids, and “Symbiont X” have apparently lost sufB altogether, which makes sense given their loss of the plastidial Fe–S cluster biosynthesis pathway and the convergently similar situation in the piroplasms (Janouškovec et al. 2019; Mathur et al. 2019). The alternative “essential tRNAs” hypothesis suggests that plastid-encoded tRNA-fMet is essential for protein synthesis of mitochondrial-encoded cytochrome oxidases (Barbrook et al. 2006). This hypothesis is consistent with the recent discovery that eugregarines have evolved reduced mitochondria known as mitochondrial-related organelles (MROs) (Salomaki et al. 2021; Mathur, Wakeman, et al. 2021). These organelles have lost both respiratory function and the mitochondrial genome, therefore eliminating their requirement of tRNA-fMet and thus allowing for the loss of the plastid genome but retention of the plastid compartment in these organisms. However, methionyl-tRNA formyltransferase is not conserved across all apicoplast genomes (such as in the piroplasms and marosporidians), suggesting it may not be the primary mechanism for plastid genome retention. Altogether, the truth might prove
to be a mix of all these explanations. The "limited transfer window" hypothesis presents a simple scenario for distribution of plastid genomes in the ACS clade due to plastid genomes being dispensable if photosynthesis is lost and if supB and clpC are successfully relocated to the nucleus or lost as well. However, the co-occurrence of MROs and plastid genome loss suggests functional interdependence between organelles may also play a role in some lineages.

The Diversity of Cryptic Plastids

The last common ancestor of the ACS clade possessed a photosynthetic plastid, and as multiple ACS lineages independently transitioned into an obligate symbiotic or parasitic lifestyle, photosynthesis was repeatedly lost and cryptic plastids were retained in parallel in several lineages. By reconstructing the proteomes of these plastids, we find parallels in the overall metabolic pathways present, yet also find functionally significant differences in the degrees of reduction in specific lineages. By clustering taxa based on their plastid proteomes, a Principal Component Analysis (PCA) reveals three distinct clusters: taxa with apicoplasts containing plastid genomes, apicoplasts without genomes, and taxa that have lost the apicoplast (fig. 3A) (refer to supplementary table S3, Supplementary Material online for a complete list of species in each cluster). Principal component 1 (x-axis) is likely reflecting the number of proteins present and accounts for nearly half of the variation in the plastid proteomes in contrast to component 2 (y-axis), which accounts for variation in the specific proteins present in the less reduced apicoplasts, as exemplified by Toxoplasma and Neospora which possess a number of lineage-specific genes. This pattern suggests a gene-loss-based convergence and reveals that the spectrum of possible outcomes for cryptic plastids once photosynthesis is lost is limited to a few main functional categories rather than a continuum of reduced plastid types. We have modeled three plastid types in figure 3B (refer to supplementary table S3, Supplementary Material online for the list of species represented by each variant). The first, represented by Toxoplasma and also includes diverse taxa such Selenidium and Digyalum (variant A), is the

![Fig. 3. The main evolutionary outcomes of cryptic plastids in the ACS clade (A) PCA species clustering based on the presence/absence of 162 apicoplast genes assigned by hyperLOPIT in Toxoplasma. Complete list of species in each cluster can be found in supplementary table S3.](https://doi.org/10.1093/molbev/msad002)

**A** PCA species clustering based on the presence/absence of 162 apicoplast genes assigned by hyperLOPIT in Toxoplasma. Complete list of species in each cluster can be found in supplementary table S3, Supplementary Material online. (B) Schematic showing the simplified metabolism of the three principal plastid variants found in the ACS clade. Black circles represent proteins. Abbreviations: MEP/DOXP, mevalonate-independent pathway/1-deoxy-D-xylulose 5-phosphate pathway; DMAPP, dimethylallyl pyrophosphate; IPP, isopentenyl pyrophosphate; FASII, fatty acid synthesis; FNR, ferredoxin NADP + oxidoreductase; Fd, Ferredoxin; Fe–S, Iron–sulfur clusters; ABC, ATP-binding cassette transporters; Tic, translocon on the inner chloroplast membrane; Toc, translocon on the outer chloroplast membrane; PUBL, plastid ubiquitin-like protein; CDC48, cell division protein; FtsH1, ATP-dependent proteases; Der1, ER-associated protein degradation (ERAD) membrane protein; Atrx, apicoplast thioredoxin protein; UFD1, ubiquitin fusion degradation protein; MRO, mitochondrion-related organelle.
one most often attributed to the apicoplast and contains all known apicoplast metabolic pathways, such as the biosynthesis of isoprenoids, Fe–S clusters, fatty acids, and heme (together with the mitochondria). It also has a suite of proteins associated with transcription, translation, and maintenance of the plastid genome. The second is represented by Platypromotum (variant B) where the plastid genome and at least one core metabolic pathway are lost (such as the loss of FASII in Platypromotum). Finally, variant C is the most extreme level of plastid reduction, represented by Lecudina, where only a single metabolic pathway remains and the plastid genome is also lost.

Interestingly, while reduction and compaction appear to be the overarching theme in cryptic plastids, lineage- and species-specific nefunctionalizations, and innovations are evident in our analyses. These proteins might be imparting novel biological functions and metabolic processes that are yet to be discovered since many are functionally unknown. This also highlights an important limitation of the data: because it is based only on proteins found in the Taxoplasma apicoplast proteome, it cannot identify proteins that are unique to other apicoplasts. In particular, an accurate inventory of proteins targeted to the plastids of photosynthetic chromerids (Chromera and Vitrella) would be highly informative to our analyses as they contain the most metabolically rich plastids in the ACS clade. However, the protein composition of the chromerid plastids is poorly known except for a recent investigation into Chromera’s photosystem proteins (Sobotka et al. 2017).

There has also been considerable progress in identifying organellar proteins in Plasmodium with proximity-dependent biotin identification and other proximity-based proteomics with the aim of discovering new antimalarial drug targets (Boucher et al. 2018; van Esveld et al. 2021). However, an important next step is the development of inducible expression systems, high-throughput proteomic techniques, and genetic studies in diverse members of the ACS group, not just the medically important parasites, to supplement homology-based studies. This would help address many of the fundamental questions that still remain elusive about these organelles such as the identity of apicoplast metabolite transporters that are vital to sustain plastidial metabolic pathways and yet remain largely unknown (Kloehn et al. 2021). Overall, the ACS clade remains the best system to model the evolution of cryptic plastid diversity and developing a better understanding of uncharacterized and novel apicoplast proteins will undoubtedly provide new insights into the principles underlying the retention of secondarily non-photosynthetic plastids and the processes of plastid reduction and loss.

Materials and Methods

Sample Collection and Sequencing

Polyporicum lacrimae was isolated from the gut of the polychaete, Notomastus tenuis, collected from the rocky pools of Grappler Inlet near the Bamfield Marine Sciences Centre, Vancouver Island, B.C., Canada (48.833 05977978153, −125.1218950544139). Cephaloidophora sp. was isolated from the gut of goose-neck barnacles (Pollicipes polymerus) collected from rocky tidepools on Brady Beach (48.829839, −125.151641) near the Bamfield Marine Sciences Centre, Vancouver Island, B.C., Canada. Selenidium orientale and S. pisinus were both isolated from the intestines of the peanut worm, Phascolosoma agassizii, collected from Ogden Point, Victoria, B.C., Canada, from sediments at a depth of 7–10 m (48.412614, −123.388713). Selenidium terebellae was collected from Clover Point, Victoria, B.C., Canada, from the intestines of the spaghett worm, Thelepus japonicus (48.40675149126726, −123.3561889595139). All specimens were collected in January 2019. Refer to supplementary figure S4, Supplementary Material online for micrographs of the sampled parasites.

The guts of the hosts were dissected using fine-tipped forceps, and hand-drawn glass pipettes were used to isolate individual gregarines (in the trophozoite life stage) under an inverted microscope. The 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 generated from pools of 2–5 cells using the Smart-Seq2 protocol (Picelli et al. 2014), and cDNA concentration was quantified on a Qubit 2.0 Fluorometer (Thermo Fisher Scientific Inc.). Sequencing libraries were prepared using the Illumina Nextera XT protocol and sequenced using the paired-end 250 bp reads Illumina MiSeq sequencer.

Transcriptome Assembly

The raw reads were trimmed using cutadapt v2.10 to remove low-quality reads, adapter sequences, and primers (Martin 2011). The remaining reads were assembled using rnaSPAdes v3.13.0 (Bushmanova et al. 2019). The resulting contigs were filtered for host contamination using BlobTools, together with BLASTn and BLASTx searches against the NCBI nt and SWISS-PROT databases (Altschul et al. 1990; Laetsch and Blaxter 2017; Bateman et al. 2021; Sayers et al. 2022). Protein-coding region prediction and translation were done using a combination of TransDecoder v5.5.0 and similarity searches against the SWISS-PROT database (Haas et al. 2013; Bateman et al. 2021). Assessment of the completeness of all transcripts generated in this study (and other transcriptomes and genomes used in further analyses) was done in BUSCO v4.0.6 with the alveolate gene marker set (Simão et al. 2015).

Phylogenomic Tree Construction and Analyses

Predicted proteins from publicly available apicomplexan genomes and transcriptomes were added to the PhyloFisher dataset for ortholog identification (Tice et al. 2021) (supplementary table S1, Supplementary Material online). Using PhyloFisher, the script fisher.py was used to identify potential orthologs from all newly added
taxa, and working_dataset_constructor.py compiled single gene FASTA files for all potential orthologs and paralogs for the newly added data and taxa spanning the eukaryotic tree of life in the PhyloFisher database (Tice et al. 2021). Each individual gene was filtered for sequencing errors and non-homologous sites using PREQUAL (Whelan et al. 2018) aligned using MAFFT (-globalpair -maxiterate 1,000) (Katoh and Standley 2013), alignment uncertainty and errors were filtered using DIVVIER (-partial -mincol 4 -divvygap) (Ali et al. 2019), and the filtered alignments were trimmed of sites comprised of >80% gaps using trimAL (-gt 0.2) (Capella-Gutiérrez et al. 2009). Trees were constructed from the trimmed alignments using IQ-TREE under the LG + C20 + F + G model, and 100 bootstrap replicates for each gene were estimated using RAxML (PROTGAMMALG4X model) (Stamatakis 2014; Nguyen et al. 2015). Bootstrap support was mapped to each tree and the resulting trees were then manually inspected to identify orthologs, paralogs, and contamination for each taxon.

Taxon selection for the final phylogenomic dataset was based on an 40% threshold for taxon occupancy in genes. For example, Toxoplasma is present in 208 of the 240 (86.67%) genes in the PhyloFisher database and is therefore included. Subsequently, individual genes that contained at least 80% of the ingroup taxa were selected for inclusion resulting in a final phylogenomic dataset containing 203 genes. These genes were filtered for sequencing errors and non-homologous sites using PREQUAL (Whelan et al. 2018), aligned using MAFFT (-globalpair -maxiterate 1,000) (Katoh and Standley 2013), alignment uncertainty and errors were filtered using DIVVIER (-partial -mincol 4 -divvygap) (Ali et al. 2019), and the filtered alignments were trimmed of sites comprised of >20% gaps using trimAL (-gt 0.8) (Capella-Gutiérrez et al. 2009). The resulting alignments were concatenated to create our 54,661-site phylogenomic dataset.

This phylogenomic dataset was subjected to ML analysis in IQ-TREE v.1.6.12 (LG + C60 + F + G) (Minh et al. 2020) with support estimated using 1,000 PMSF bootstrap replicates. Bayesian inferences were conducted in Phylobayes (Lartillot et al. 2009) under the CAT-GTR model with four chains for each dataset run for >20,000 generations each, after which the chains were nearing convergence on the same topology (burnin = 4,000; maxdiff = 0.30). Single gene datasets including all considered sequences [i.e., including orthologs and paralogs; see (Salomaki et al. 2020)] and ortholog-only datasets are available in Mendeley Data, V1, doi: 10.17632/fps2p4m25d.1.

Removal of Fast Evolving and Heterotachous Sites, and Random Gene Subsampling

Scripts within the PhyloFisher software package were used to construct datasets for fast-evolving site removal, heterotachous site removal, and random gene subsampling (Tice et al. 2021). The fastest evolving sites were sequentially removed in 3,000 site chunks generating new alternative datasets at each step until all sites were removed. Similarly, the most heterotachous sites were removed in a stepwise fashion, 3,000 sites at a time, producing iteratively smaller datasets until no further sites could be removed. Genes from our 99 taxa datasets were randomly subsampled in sets of 20% (n = 14), 40% (n = 6), 60% (n = 4), and 80% (n = 2) of the complete dataset, under the default 95% confidence interval setting as in Salomaki et al. (2021). For all datasets generated by these scripts, 100 rapid bootstrap replicates were generated using RAxML under the LG4X model (Stamatakis 2014). Bipartitions for clades of interest were calculated using bipartition_examiner.py in PhyloFisher, and the data were plotted using ggplot in R to show the impact of these analyses on the support for relationships of interest (Tice et al. 2021).

Identifying Plastid Proteins Across the ACS Clade

Curating the Proteome Dataset

The hyperOPIT spatial proteomics data, specifically the 162 T. gondii ME-49 proteins assigned to the apicoplast, were acquired from the supplementary tables provided by Baryllyk et al. (2020). To search in-depth for homologs of putative apicoplast proteins across the ACS clade, an in-house curated set of 75 proteomes was used. This dataset includes newly generated transcriptomes from this study, multiple representative species from all major apicomplexan subgroups, and all publicly available sequencing data for the gregarines, chrompodellids, and squirmids. Ciliate and opisthokont taxa were also downloaded from the EukProt database and included in our dataset as outgroups to help with the identification of putative paralogs in downstream analyses (Richter et al. 2022). Species with BUSCO completeness scores <40% were ultimately removed from the final analysis as the low completeness may indicate that a putative orthologue is not necessarily absent but missing due to an incomplete sequencing dataset (supplementary table S1, Supplementary Material online).

Searching for Orthologs

For each of the 168 putative apicoplast proteins in T. gondii ME-49, a Profile HMMs search against our proteome dataset was carried out using HMMR v3.1 (e-value 1e-5) and BLASTp (e-value 1e-25) (Altschul et al. 1990; Eddy 2011). The resulting hits were extracted, aligned, and trimmed, using MAFFT v7.222 (-auto option) and trimAl v1.2 (-gt 0.2), respectively (Capella-Gutiérrez et al. 2009; Katoh and Standley 2013), and used to generate phylogenies in FastTree v2.1.3 (Price et al. 2010). The individual trees were manually evaluated in FigTree, and contaminants, deep paralogs, and long-branching divergent sequences were identified and removed (Rambaut 2014). The remaining sequences were realigned and used to generate ML phylogenies in IQ-TREE v.1.6.12 (Minh et al. 2020). 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 bootstraps (Kalyaanamoorthy et al. 2017). The presence of targeting signals in putative apicoplast proteins was assessed using SignalP v6.0 and TargetP v2.0 (Armenteros et al. 2019; Teufel et al. 2022). The resulting putative ortholog sequences for each protein query can be found in FASTA files (Data S1). The homologs were also used as queries for BLASTp searches against the NCBI nr database, and homologs that recovered the specific Toxoplasma protein within their best matches were retained (Altschul et al. 1990). Those that did not were further scrutinized to ensure they branched in the correct position, and were potentially divergent homologs.

Assigning Functional Annotation
The putative apicoplast proteins were given functional annotations with eggNOG, Pfam, and KEGG, using a combination of eggNOG-Mapper v2 and downloading annotations of the T. gondii ME-49 proteins from UniProt (Kanehisa et al. 2016; Huerta-Cepas et al. 2019; Bateman et al. 2021; Mistry et al. 2021). Ultimately 123 of the proteins received an annotation, while those that did not have any GO term, Pfam, or KO term assigned. We must note that some proteins were annotated as containing “DUF” domains; however, the function of these protein domains remains to be elucidated.

Data Visualization
The binary patterns of our putative orthology search were used to generate the presence/absence plot in figure 2 using the tree annotation tool in iTOL v6.5.7 (Letunic and Bork 2021). The PCA plot in figure 3 was generated using the “ggbiplot” package in R (http://github.com/vqv/ggbiplot).

Supplementary Material
Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments
We are grateful to Nick Irwin for assistance with sampling and data processing in R. This work was supported by a grant from the Gordon and Betty Moore Foundation to P.J.K. (https://doi.org/10.37807/GBMF9201). V.M. was supported by the Li Tze Fong affiliated graduate fellowship from the University of British Columbia. E.D.S. was supported by the MSCA-IF SMART (CZ.02.2.69/0.0/0.0/ 20_079/0017809). M.K. was supported by ERD fund “Centre for Research of Pathogenicity and Virulence of Parasites” (no. CZ.02.1.01/0.0/0.0/16_019/0000759). We would also like to thank to IT4Innovations National Super Computer Center, Technical University of Ostrava, Ostrava, Czech Republic for access to computational resources (projects #Open-24-68).

Data Availability
RNA-seq raw reads are available in the NCBI SRA under accession number PRJNA872092. Transcriptome assemblies, phylogenomics matrices, and individual protein alignments are available at Mendeley Data, V1, doi: 10.17632/fps2pmn25d.1.

References


