Phylogenomics Identifies a New Major Subgroup of Apicomplexans, Marosporida class nov., with Extreme Apicoplast Genome Reduction

Varsha Mathur1,*, Waldan K. Kwong1, Filip Husnik 2, Nicholas A.T. Inrin 1, Árni Kristmundsson3, Camino Gestal4, Mark Freeman5, and Patrick J. Keeling1

1Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada
2Okinawa Institute of Science and Technology, Okinawa, Japan
3Fish Disease Laboratory, Institute for Experimental Pathology, University of Iceland, Reykjavik, Iceland
4Institute of Marine Research (IIM-CSIC), Vigo, Spain
5Ross University School of Veterinary Medicine, Basseterre, Saint Kitts and Nevis, West Indies

*Corresponding author: E-mail: varsha.mathur@botany.ubc.ca.
Accepted: 17 November 2020

Abstract

The phylum Apicomplexa consists largely of obligate animal parasites that include the causative agents of human diseases such as malaria. Apicomplexans have also emerged as models to study the evolution of nonphotosynthetic plastids, as they contain a relict chloroplast known as the apicoplast. The apicoplast offers important clues into how apicomplexan parasites evolved from free-living ancestors and can provide insights into reductive organelle evolution. Here, we sequenced the transcriptomes and apicoplast genomes of three deep-branching apicomplexans, Margolisiella islandica, Aggregata octopiana, and Merocystis kathae. Phylogenomic analyses show that these taxa, together with Rhytidocystis, form a new lineage 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 plastid metabolism. However, the apicoplast genomes of Margolisiella and Rhytidocystis are the most reduced of any apicoplast, are extremely GC-poor, and have even lost genes for the canonical plastidial RNA polymerase. This new lineage of apicomplexans, for which we propose the class Marosporida class nov., occupies a key intermediate position in the apicomplexan phylogeny, and adds a new complexity to the models of stepwise reductive evolution of genome structure and organelle function in these parasites.

Key words: organelle evolution, plastids, apicomplexans, phylogenomics.

Significance

Apicomplexans are obligate parasites of animals, however, they evolved from algae and retain a nonphotosynthetic plastid (the apicoplast). Using phylogenomics, we resolve the branching order of major apicomplexan lineages, and identify a new class of apicomplexans, the Marosporida. We also show marosporidians have the most reduced apicoplast genomes sequenced to date, which lack canonical plastidial RNA polymerase and provide new insights into reductive organelle evolution.

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
**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 (Janouškové 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, nonphotosynthetic chloroplast, known as the apicoplast. Since its surprising discovery (McFadden et al. 1996; Wilson 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–sulfur clusters (Lim and McFadden 2010). Although 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-plastid 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 *Margolišiella* 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.

**Results and Discussion**

A Resolved Multiprotein Phylogeny of the Apicomplexa

We generated new transcriptomes and WGS sequencing data for *M. kathae* and *Ma. islandica*, and WGS data from *A. octopiana*, after isolating the parasites from their marine mollusc hosts (SRA PRJNA645464). *Margolišiella islandica* is known to infect Icelandic scallops (*Chlamys islandica*) where it causes an intracellular infection in the heart auricle (Kristmundsson et al. 2011). *A. 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, 2019), and *M. kathae* infects the renal tissues of the common whelk (*Buccinum undatum*) with intermediate life stages in scallops (Kristmundsson and Freeman 2018) (fig. 1Aa–c). 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.

To place these species within a phylogenomic context, we added them to a data set of slow-evolving nuclear genes previously used to resolve deep phylogenetic relationships within the Apicomplexa (Mathur et al. 2019). This data set 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 *Margolišiella islandica*. Using phylogenomics, we present a robust multigene apicomplexan phylogeny incorporating all published apicomplexan taxa to date (Janouškové et al. 2019; Mathur et al. 2019; Muñoz-Gómez et al. 2019). We show that species of the genera *Aggregata*, *Merocystis*, *Margolišiella*, 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 present a robust multigene apicomplexan phylogeny incorporating all published apicomplexan taxa to date (Janouškové et al. 2019; Mathur et al. 2019; Muñoz-Gómez et al. 2019). We show that species of the genera *Aggregata*, *Merocystis*, *Margolišiella*, 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 *Margolišiella* 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.
**Fig. 1.**—Phylogeny of the Apicomplexa. (A) Light micrographs of oocysts of the species sequenced, left to right, *Merocystis kathae*, *Margolišea islandica*, and *Aggregata octopiana*. Scale bars are indicated on the figure. (B) 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 nonparametric bootstrap support (1,000 replicates, LG+F+C19 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 S1, Supplementary Material online. * denotes that Nephromyces is a chimeric OTU assembled from several most closely related lineages from inside the renal sac of a single host. (C) 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 nonparametric replicates (IQ-TREE model LG+F+C19). * denotes taxa that only have plastid genome data available.

was also expanded to incorporate 11 other recently published transcriptomes (Janouškové 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 (supplementary table S1, Supplementary Material online). Maximum-likelihood phylogenies using an empirical profile mixture model (LG+C40+F+C4) 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 (fig. 1B). 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 *Margolišea islandica* all branch with *Rhityocystis* in a robustly supported monophyletic group (fig. 1B). 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) (supplementary fig. S1, Supplementary Material online). However, the Aggregatidae has been placed within the Coccidia, which is not consistent with the phylogenomic tree (fig. 1B). Similarly, the placement...
of Margoliisiella as the sister taxon to Rhytidocystis has also been observed previously in rRNA phylogenies, although with variable statistical support (Kristmundsson et al. 2011; Mirolubova et al. 2020) (supplementary fig. S1, Supplementary Material online). Historically, however, Rhytidocystis and Margoliisiella 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 Agregata or Merocystis (Mirolubova et al. 2020). Here, we show with robust multiprotein phylogenomics that these taxa are sisters, but are not coccidians (fig. 18).

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 confusion 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 Aggregata is therefore transferred from the Coccidia to the Marosporida to reflect the sister relationship of Agregata and Merocystis, and similarly the Rhytidocystidae, which was erected for the genus Rhytidocystis (Levine 1979; Leander and Ramey 2006; Rueckert and Leander 2009), can also be transferred to the Marosporida. The Eococcidia (Mirolubova 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; Mirolubova et al. 2020). Whether the corallicolids are also members of Marosporida is still an open question. Corallicolid transcriptomic data remain unavailable, however, 18S rRNA and mitochondrial gene phylogenies do not support this grouping (supplementary figs. S1 and S2, Supplementary Material online; Mirolubova 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.

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 apicomplexan genomes are poorly sampled, making phylogenetic reconstructions less reliable. To fill this gap, we reanalyzed the apicoplast phylogeny with significantly greater taxonomic diversity. To obtain apicoplast genomes from Agregata, Merocystis, and Margoliisiella, we conducted 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, Siedlecki, Selendium, 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 (supplementary table S2, Supplementary Material online). Using this matrix in combination with ML and Bayesian phylogenetic analyses, we recovered a poorly resolved phylogeny (fig. 1C).

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 (fig. 1Q). Both phylogenies fully support the sister relationships between Merocystis and Agregata, and Margoliisiella and Rhytidocystis. The plastid phylogeny also recovers a monophyletic grouping of the Hematozoa, Piroplasmida, and Coccidia. However, the positions of the gregarines, Selendium and Siedlecki, as well as Nepromycoses, 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 (supplementary fig. S3, Supplementary Material online), 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 (supplementary fig. S4, Supplementary Material online). The node placing Agregata 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 artifacts. 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 with 58,611 sites in the nuclear tree), we conclude, in agreement with the findings of Muñoz-Gómez et al. (2019), that the apicoplast-based analyses are less robust in resolving phylogenetic relationships within the Apicomplexa.

**Apicoplast Genomes in Margoliisiella 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 (fig. 2A).
They contain compact (~38 kb) 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 (chlA, chlB, chlC, and acsF).

The only significant differences between the two genomes are the presence of the ribosomal protein gene rpl11 and the RNA polymerase (RNAP) 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 (fig. 2B).

Unlike Merocystis and Aggregata, the plastid genome of Margolisiella is strikingly reduced compared with all other apicoplasts sequenced to date (fig. 2A and B). The genome is very small (18 kb), 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 (fig. 2A 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, Nephromyces, and the two chromodendellids, Piridium and Chromera (Janouškovec et al. 2010; Kwong et al. 2019; Mathur et al. 2019; Muñoz-Gómez et al. 2019). Unlike all other known apicoplasts, the Margolisiella apicoplast genome has lost all four of its plastid-encoded RNAP genes, which are presumed to be solely responsible for the transcription of the apicoplast genome and therefore functionally indispensable (Nisbet and McKenzie 2016). A similarly AT-rich fraction of sequence reads was also observed by Janouškovec et al. (2019) in two species of Rhytidocystis, where the authors found apicoplast proteins in their transcriptome data. 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 forbesi) 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 S. nematoides and E. duboscqi (supplementary fig. S5, Supplementary Material online) and fragmented genomic contigs that included most of the expected genes from Selenidium and Dicygalum (see Materials and Methods) (supplementary table S2, Supplementary Material online).

The Rhytidocystis apicoplast genomes are even more reduced than Margolisiella (13 and 14 kb, in Rhytidocystis sp. 1 and sp. 2, 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 (Su et al. 2019). 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 (fig. 2B). Strangely, Rhytidocystis sp. 1 uses an alternate genetic code (UGA encodes tryptophan), but Rhytidocystis sp. 2 uses the standard genetic code (fig. 2B). Both genomes lack all genes for RNAP, 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 RNAPs in Margolisiella and Rhytidocystis raises the question of how their apicoplast genes are transcribed. We first explored the possibility that the RNAP 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 RNAPs 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 RNAP 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 (fig. 3A). 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 RNAP has been lost entirely, and apicoplast transcription relies on a separate and distinct nuclear-encoded polymerase derived from some other...
proteins were found. (RNAPs I, II, and III and other RNAPs with uncertain phylogenetic associa-
tions). Numbers within the circles denote the number of unique proteins
attached to the rest of the protein, and no line indicates fragmented
circles signifies a complete T7 polymerase where the N-terminal domain is
proteins. Genes, and the ancestral plastid polymerase that transcribes
from T7 bacteriophage that transcribes nonphotosynthesis
photosynthesis-related genes (Liere et al. 2011). Intriguingly,
plastid-encoded RNAP, and in the genus
Margolisiella
new major subgroup of apicomplexans, Eleutheroschizon and Piridium (fig. 3B). The presence of eukaryotic
RNAPs I, II, and III and other RNAPs with uncertain phylogenetic associa-
tion, 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.

source. Land plant plastids, for example, use two different
RNAPs: a nuclear-encoded polymerase related to homologs
from T7 bacteriophage that transcribes nonphotosynthesis-related 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 RNAP, 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 transcripts 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
RNAP subunits, phage-type RNAPs, and other viral RNAPs (El-
Gebali et al. 2018) (fig. 3A). We identified subunits of the
eukaryotic RNA polymerases (RNAPI, II, and III) in all the api-
complexans searched, as well as a few phylogenetically am-
biguous proteins that were not associated distinctly with a
particular polymerase (labeled as “uncertain”) (fig. 3A).

These searches also found T7 bacteriophage derived
RNAPs in all taxa, except for Merocystis (fig. 3A), which may
be due to the incompleteness of the Merocystis transcriptome
(refer to supplementary table S1, Supplementary Material online, for BUSCO completeness scores). Most eukaryotic mito-
chondria 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 mitochondrionally targeted polymerases. In Selendium, Siedleckia, and
Eleutheroschizon we retrieved truncated proteins that lack
the complete N-terminal domain, whereas in Toxoplasma,
Piridium, and Dicyugales, we recovered the complete protein
(fig. 3B). Intriguingly, we found additional T7 polymerases in several apicomplexans, including Rhytidocystis. Two N-termi-
mal domain fragments were found in Rhytidocystis sp. 1, whereas in Rhytidocystis sp. 2, we found two truncated T7
polymerase transcripts that are missing the N-terminal domain
(fig. 3B). Margolisiella contained two nonoverlapping T7 poly-
merase 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 apico-
plast, or even that the mitochondrial T7 polymerase is dually
targeted in these nonmodel species. A precedent comes from
land plants, where a dually targeted RNA with an ambiguous
targeting sequence allows a mitochondrial T7 polymerase to
be imported into both the chloroplast and the 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 alterna-
tive 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 unrec-
ognized but ancient redundancy in apicomplexan RNAPs
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 Margolisiella. Generally, apicoplasts are in-
volved in biosynthesis of isoprenoids (MEP), fatty acids
(FASII), iron–sulfur (Fe–S) clusters, and part of the tetrapyrrole
(heme) biosynthesis pathway (Sheiner et al. 2013). Most api-
complexans retain genes for all four pathways, however, the
piroplasms and “Symbiont-X” only carry out isoprenoid bio-
synthesis (Janouškovcov et al. 2015, 2019), and the marine
gregarine clade that includes Pterospora, Lankesteria, and
**Summary and Future Directions**

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 metabolism of this new group is conserved, but the apicoplast genome structure and content are highly variable. Apicoplasts from *Margoliussiella* and *Rhytidocystis* have the smallest, most reduced, and most AT-biased genomes known to date, and have distinctively lost plastid-encoded RNAP and retention and loss in deep-branching apicomplexans is more complex than previously thought, as are other relatively stable characters, such as noncanonical 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 squirrels. We also still lack nuclear genomic resources for several important groups, such as coralciliates 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.

**Materials and Methods**

**Sample Collection and DNA/RNA Extraction**

*Meroeystis kathae* was isolated from a common whelk (*B. undatum*) and *Ma. islandica* was isolated from an Iceland scallop (*C. 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% MgSO4 in seawater for 1–2 h. The renal organ of the whelks was examined under a dissecting microscope for the presence of *Meroeystis* infections. Subsequently, the relatively large gametogenic stages of *Meroeystis* 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 *Margoliussiella* (all life stages present) examined from a small drop of hemolymph from the
auricles. Small samples from heavily infected heart auricles, and its hemolymph, were taken for molecular analysis. DNA and RNA from the samples, infected with *Meroçystis* and *Margolisella*, was extracted using a QIAGEN AllPrep DNA/RNA Mini Kit (Cat. No. 80204).

*Aggregata octopiana* was isolated from a pool of five infected octopuses (*O. vulgaris*) collected using traps by local fishermen in Ria de Vigo, Spain (24°14.09’N, 8°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 analyze 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 ml of filtered seawater with 1% Tween 80 using an electric tissue grinder (IKA-Ultra Turrax T-25). Tissue homogenates were filtered twice with a nylon mesh of 100 and 41 μm, respectively, to remove tissue fragments. The filtrate was then centrifuged at 1,000 g for 5 min. The sporocysts were cleaned using a density gradient centrifugation method according to Gestal et al. (1999), and counted in a Neubauer chamber to standardize the sample at 2 × 10⁶ sporocyst/ml. Light microscopy was used to analyze the morphology and dimensions of the fresh sporocysts. For DNA extraction, sporocysts were resuspended in 500 μl of extraction buffer (NaCl 100 mM, EDTA 25 mM 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 ml⁻¹; 37°C overnight), genomic DNA was purified using a phenol:chloroform:isoamyl alcohol extraction method (Sambrook et al. 1989). DNA was precipitated with ethanol and sodium acetate overnight at −20°C. 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 WGS 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. All sequencing raw reads have been deposited under SRA PRINa645464. The quality of the raw reads was assessed using FastQC and trimmed to remove sequencing adaptors with Trimmomatic v0.36 (Bolger et al. 2014; Andrews et al. 2015). 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 and Blaxter 2017; 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 (Diercksens et al. 2016). The plastid genomes of *S. nematoides*, *Eleutheroschizon duboscpi*, and *Rhytidoctys* sp. 1 and sp. 2 were assembled based on plastid and mitochondrial overlaps from assemblies of raw transcript reads using maSPAdes 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 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 data set comprising 40 plastid-encoded proteins was compiled based on a previously published data set (Mathur et al. 2019) and enriched with all apicomplexan plastids that have been sequenced as of January 2020 (supplementary table S2, Supplementary Material online) as well as the plastid-encoded proteins of *A. octopiana*, *M. kathae*, and *M. islandica*. Profile 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 supplementary table S2, Supplementary Material online, 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 5,759 amino acid positions from 58 taxa (available at Mendeley data doi:10.17632/rrdc4xsk2h.1). The phylogenomic maximum-likelihood analyses were done in IQ-TREE.
v.1.5.4 with the model LG+F+I+R7 and 500 nonparametric bootstraps. This model best fits the data according to the Akaike information criterion and the corrected Akaike information criterion 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 nonredundant 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. All raw reads have been deposited under SRA PRJNA645464. 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 adapter and primer sequences were trimmed using Trimmomatic v0.36 and the transcriptomes was 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 NCBI nt database and the SWISS-PROT database, respectively (Agarwala et al. 2018; Bateman 2019). The contigs that contained hits were extracted and coding regions were predicted using Exonerate v2.2.0 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 v7.722 and trimmed using trimAl v0.9.6, and FastQC was used to assess the quality of the filtered alignments. The trimmed contigs were 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 (available at Mendeleev data doi:10.17632/rrdc4xsk2h.1). The phylogenomic maximum-likelihood tree was constructed with the heterogeneous mixture LG+C40+F+I+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+I+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 chains were run for 10,000 generations and two chains converged with max diff. = 0.017, whereas 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 RNAPs in *Margolisella* 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 phase-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 nonapicomplexan outgroups (Bateman 2019). Identified SWISS-PROT and apicomplexan proteins were aligned using MAFFT v7.222 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 polymersome complexes these proteins corresponded to (e.g., eukaryotic RNAPI, RNAPIII, RNAPII, or prokaryotic RNAP), proteins were annotated using BlastP searches against the SWISS-PROT database (max_target_seqs 1, E < 1e-5) and their phylogenetic context was interpreted in FigTree. Phylogenetically ambiguous proteins that were not clearly associated with a certain polymersome were labeled 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 data set, the WGS contigs of *Aggregata*, *Merocystis*, and *Margolisella* 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 v2.2.0 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. Phylogenetically ambiguous proteins were labeled as “uncertain.”

**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 protein 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 v2020.1.1. 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.

**Supplementary Material**

Supplementary data are available at Genome Biology and Evolution online.

**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 P.J.K. 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.

**Author Contributions**

V.M. and P.J.K. designed the study. A.K., M.F., and C.G. obtained samples. V.M. and F.H. performed transcriptomics and WGS preparation. V.M., F.H., and W.K.K. assembled and annotated the plastids. V.M. and N.A.T.I. carried out the RNAP analysis. V.M. analyzed the rest of the data. V.M. and P.J.K. wrote the paper with input from all authors.
Data Availability

The plastid genomes generated here are available at the NCBI GenBank Nucleotide Database (www.ncbi.nlm.nih.gov/nucleotide/) and can be accessed with the following accession numbers: MW088710, MW088711, and MW088712. The raw sequencing reads are available on the Sequence Read Archive (www.ncbi.nlm.nih.gov/sra) and can be accessed with the accession PRJNA645464. The alignments for the phylogenomics analyses and transcriptome-mined plastid genomes are available at Mendeley Data (www.mendeley.com/datasets) and can be accessed with the doi:10.17632/rddc4xsk2h.3.

Literature Cited


Associate editor: McFadden Geoff