Report

Current Biology

Multiple Independent Origins of Apicomplexan-Like Parasites

Highlights

- The origin of apicomplexans from algae occurred at least three times independently
- *Piridium* and *Platyproteum* form distinct lineages of obligate animal parasites
- They both retain cryptic plastids that are highly convergent with apicomplexans

Authors

Varsha Mathur, Martin Kolísko, Elisabeth Hehenberger, ..., Árni Kristmundsson, Mark A. Freeman, Patrick J. Keeling

Correspondence

varsha.mathur@botany.ubc.ca

In Brief

Apicomplexans are a major group of diverse animal parasites thought to have evolved once from free-living photosynthetic algae. Mathur et al. show, using single-cell genomics, that this complex evolutionary shift actually occurred at least three times independently, resulting in multiple lineages of highly convergent animal parasites.

Mathur et al., 2019, Current Biology *29*, 2936–2941
September 9, 2019 © 2019 Elsevier Ltd.
https://doi.org/10.1016/j.cub.2019.07.019

Multiple Independent Origins of Apicomplexan-Like Parasites

Varsha Mathur,[1,7,8,9,*] Martin Kolísko,[1,2,7] Elisabeth Hehenberger,[1,3] Nicholas A.T. Irwin,[1] Brian S. Leander,[1,4] Árni Kristmundsson,[5] Mark A. Freeman,[6] and Patrick J. Keeling[1]

[1]Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[2]Institute of Parasitology, Biology Centre, Czech Acad. Sci., Branišovská 31, České Budějovice 370 05, Czech Republic
[3]GEOMAR - Helmholtz Centre for Ocean Research, Duesternbrooker Weg 20, 24105 Kiel, Germany
[4]Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[5]Institute for Experimental Pathology, University of Iceland, Keldur. Keldnavegur 3, 112 Reykjavík, Iceland
[6]Ross University School of Veterinary Medicine, PO Box 334, Basseterre, St. Kitts, West Indies
[7]These authors contributed equally
[8]Twitter: @varsh_mathur
[9]Lead Contact
*Correspondence: varsha.mathur@botany.ubc.ca
https://doi.org/10.1016/j.cub.2019.07.019

SUMMARY

The apicomplexans are a group of obligate animal pathogens that include *Plasmodium* (malaria), *Toxoplasma* (toxoplasmosis), and *Cryptosporidium* (cryptosporidiosis) [1]. They 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. 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 [2]. Their closest relatives include a mix of both photosynthetic algae (chromerids) and non-photosynthetic microbial predators (colpodellids) [3]. Genomic analyses of these free-living relatives have revealed a great deal about how the alga-parasite transition may have taken place, as well as origins of parasitism more generally [4]. 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.

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 1A–1E). 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 1F). 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 [5–8] (Figure S1). *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 [9, 10]. *Piridium sociabile* is even more poorly studied: found in 1932 as an intracellular infection and was morphologically classified as a schizogregarine [11].

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 [12, 13]. Their positions in the resulting tree are strongly and consistently resolved by both maximum likelihood (C40+LG+Γ4+F model) and Bayesian (CAT-GTR) analyses (Figure 1G). 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” (chromerids + colpodellids), with complete support as sister to the photosynthetic alga *Vitrella brassicaformis*. *Platyproteum* forms a new lineage, also with complete

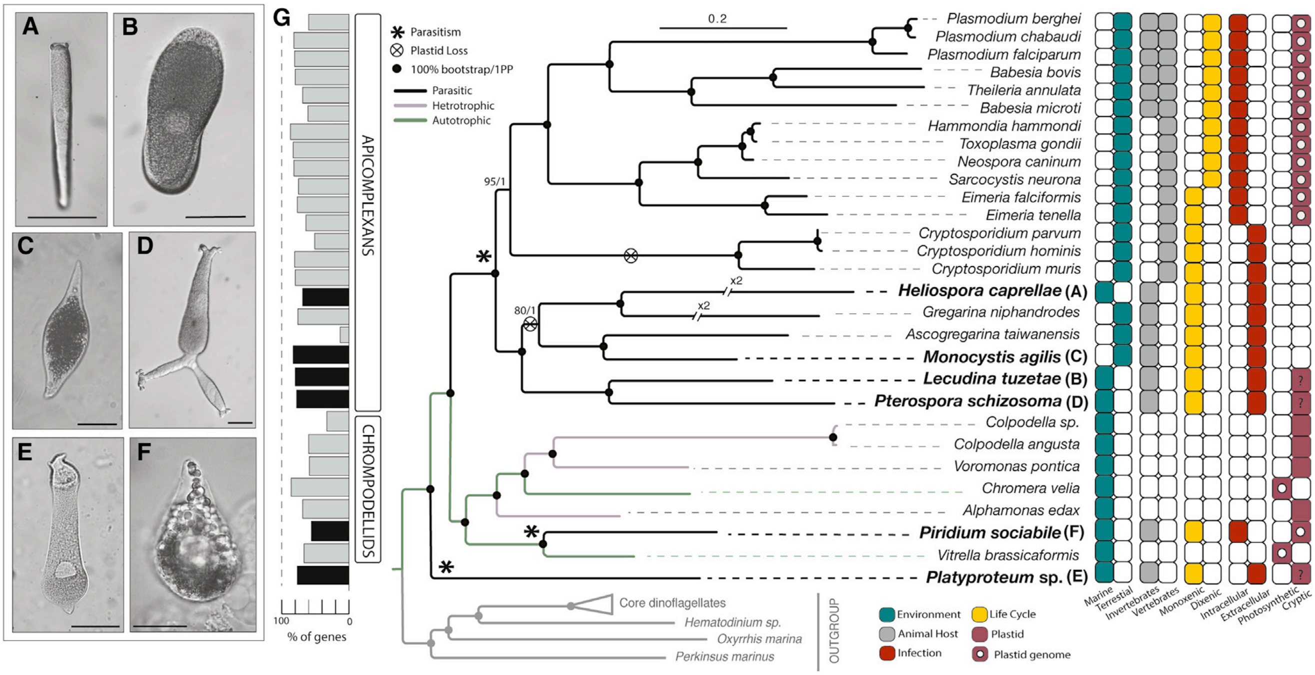

Figure 1. Phylogenomic Tree of the Apicomplexa and Relatives

(A–E) 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+Γ4+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.

See also Figures S1 and S3.

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.

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 [2, 3]. With both genomic and transcriptomic data from *Piridium*, we first assembled its complete plastid genome (Figures 2A and S2), which is strikingly similar in size, architecture, and gene content to apicoplast genomes (Figure 2B). The *Piridium* plastid genome is a highly reduced compact circle (~34 kb) 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*) [14, 15]. 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 2C). 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* [16, 17].

Apicomplexans depend on apicoplasts for essential biosynthesis of four compounds: isoprenoids (the non-mevalonate pathway); heme; iron-sulfur (Fe-S) clusters; and fatty acids (the type II fatty acid pathway) [2]. 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 [18, 19]. 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 2C; Table S1). 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

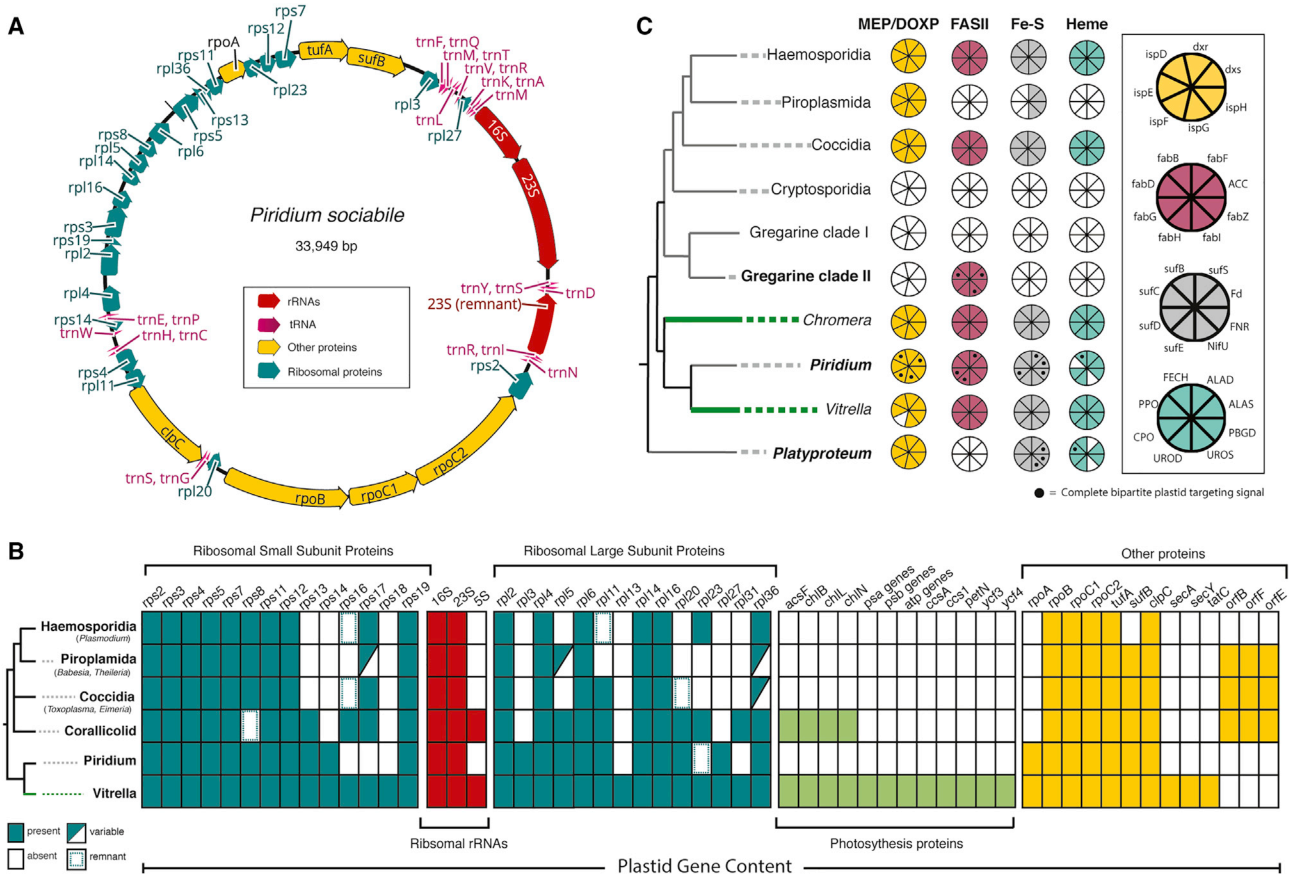
A
Piridium sociabile
33,949 bp
rRNAs
tRNA
Other proteins
Ribosomal proteins
rps7
rps12
rpoA
rps11
rpl36
tufA
sufB
rpl23
rps13
rps5
rpl6
rps8
rpl5
rpl14
rpl16
rps3
rps19
rpl2
rpl4
rps14
trnW
trnE, trnP
trnH, trnC
rps4
rpl11
clpC
trnS, trnG
rpl20
rpoB
rpoC1
rpoC2
rps2
trnR, trnI
trnN
trnY, trnS
23S (remnant)
trnD
23S
16S
rpl3
trnL
rpl27
trnF, trnQ
trnM, trnT
trnV, trnR
trnK, trnA
trnM
C
MEP/DOXP
FASII
Fe-S
Heme
Haemosporidia
Piroplasmida
Coccidia
Cryptosporidia
Gregarine clade I
Gregarine clade II
Chromera
Piridium
Vitrella
Platyproteum
ispD dxr dxs ispH ispG ispF ispE
fabB fabF ACC fabZ fabI fabH fabG fabD
sufB sufS Fd FNR NifU sufE sufD sufC
FECH ALAD ALAS PBGD UROS UROD CPO PPO
● = Complete bipartite plastid targeting signal
B
Ribosomal Small Subunit Proteins
Ribosomal Large Subunit Proteins
Other proteins
rps2 rps3 rps4 rps5 rps7 rps8 rps11 rps12 rps13 rps14 rps16 rps17 rps18 rps19
16S 23S 5S
rpl2 rpl3 rpl4 rpl5 rpl6 rpl11 rpl13 rpl14 rpl16 rpl20 rpl23 rpl27 rpl31 rpl36
acsF chlB chlL chlN psa genes psb genes atp genes ccsA ccs1 petN ycf3 ycf4
rpoA rpoB rpoC1 rpoC2 tufA sufB clpC secA secY tatC orfB orfF orfE
Haemosporidia (Plasmodium)
Piroplamida (Babesia, Theileria)
Coccidia (Toxoplasma, Eimeria)
Corallicolid
Piridium
Vitrella
present
variable
absent
remnant
Ribsomal rRNAs
Photosynthesis proteins
Plastid Gene Content

Figure 2. 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*.

See also Table S1 and Figure S2.

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 [17, 18].

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) [19, 20]. 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 [3]. The gregarines thus suggest that plastid dependency is highly context dependent.

Looking beyond the plastid, metabolic reconstructions based on KEGG (Kyoto Encylopedia of Genes and Genomes) identifiers across the whole genome confirm an overall convergence of functional reduction but also some divergence (Figure 3). 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 S2). 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.

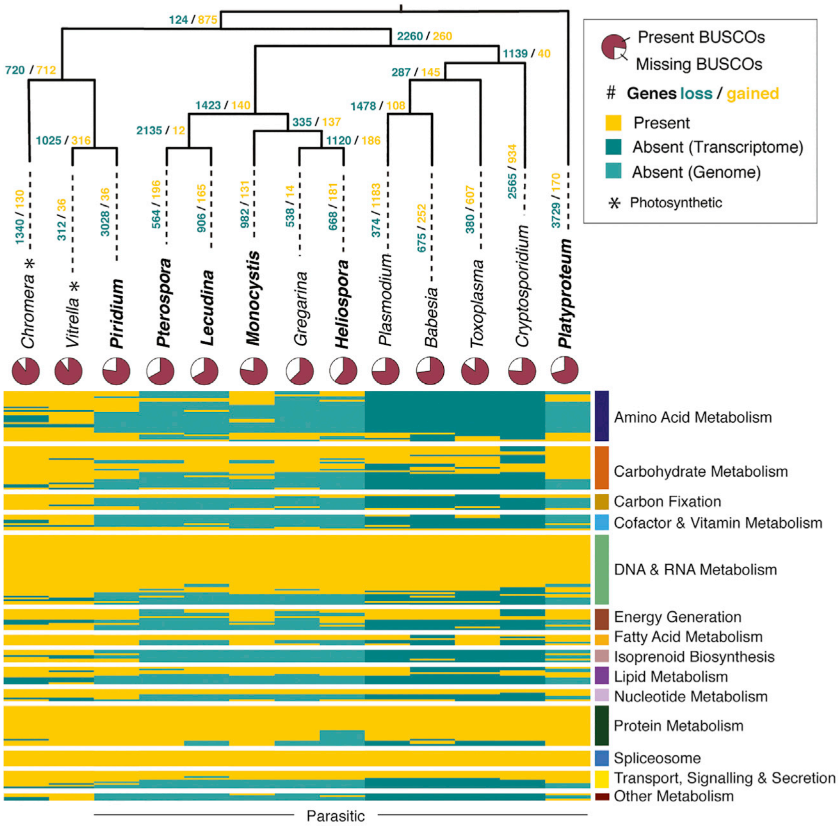

Figure 3. The Distribution of Cellular Metabolic Pathways across the Tree of Apicomplexans and Chrompodellids

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.
See also Table S2.

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.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - ○ Genomics and transcriptomics of Piridium sociabile
  - ○ Genome assembly and annotation of Piridium sociabile
  - ○ Transcriptomics of the gregarines and Platyproteum
  - ○ Transcriptome assembly and annotation of the gregarines and Platyproteum
  - ○ Ortholog identification, gene concatenation and phylogenomics
  - ○ Search and identification of plastid proteins
  - ○ Search for plastid localization signals
  - ○ Analysis of cellular metabolic pathways
  - ○ Ortholog identification and search for apicomplexan invasion/extracellular proteins
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2019.07.019.

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 Fellowship Purkyne, Czech Acad. Sci. N.A.T.I. was supported by an NSERC Canadian Graduate Scholarship.

AUTHOR CONTRIBUTIONS

V.M., M.K., and P.J.K. designed the study. V.M., Á.K., M.A.F., and B.S.L. obtained samples. M.A.F. and Á.K. performed the initial molecular diagnosis for *Piridium*. V.M., M.K., and N.A.T.I. conducted genomics, transcriptomics, and phylogenomics. E.H., N.A.T.I., M.K., and V.M. performed plastid analyses. V.M. and M.K. performed all other analyses. V.M., M.K., and P.J.K. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 22, 2019
Revised: May 28, 2019
Accepted: July 8, 2019
Published: August 15, 2019

REFERENCES

1. Votýpka, J., Modrý, D., Oborník, M., Šlapeta, J., Lukeš, J., and Votýpka, J. (2017). Apicomplexa. In Handbook of the Protists, J.M. Archibald, A.G.B. Simpson, and C.H. Slamovits, eds. (Springer International Publishing), pp. 567–624.
2. McFadden, G.I., and Yeh, E. (2017). The apicoplast: now you see it, now you don't. Int. J. Parasitol. *47*, 137–144.
3. Janouškovec, J., Tikhonenkov, D.V., Burki, F., Howe, A.T., Kolísko, M., Mylnikov, A.P., and Keeling, P.J. (2015). Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. Proc. Natl. Acad. Sci. USA *112*, 10200–10207.
4. Woo, Y.H., Ansari, H., Otto, T.D., Klinger, C.M., Kolísko, M., Michálek, J., Saxena, A., Shanmugam, D., Tayyrov, A., Veluchamy, A., et al. (2015). Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. eLife *4*, e06974.
5. Leander, B.S., Lloyd, S.A.J., Marshall, W., and Landers, S.C. (2006). Phylogeny of marine Gregarines (Apicomplexa)–*Pterospora, Lithocystis* and *Lankesteria*–and the origin(s) of coelomic parasitism. Protist *157*, 45–60.
6. Rueckert, S., Villette, P.M.A.H., and Leander, B.S. (2011). Species boundaries in gregarine apicomplexan parasites: a case study-comparison of morphometric and molecular variability in *Lecudina* cf. *tuzetae* (Eugregarinorida, Lecudinidae). J. Eukaryot. Microbiol. *58*, 275–283.
7. Rueckert, S., Simdyanov, T.G., Aleoshin, V.V., and Leander, B.S. (2011). Identification of a divergent environmental DNA sequence clade using the phylogeny of gregarine parasites (Apicomplexa) from crustacean hosts. PLoS ONE *6*, e18163.
8. Leander, B.S., Clopton, R.E., and Keeling, P.J. (2003). Phylogeny of gregarines (Apicomplexa) as inferred from small-subunit rDNA and β-tubulin. Int. J. Syst. Evol. Microbiol. *53*, 345–354.
9. Rueckert, S., and Leander, B.S. (2009). Molecular phylogeny and surface morphology of marine archigregarines (Apicomplexa), *Selenidium* spp., *Filipodium phascolosomae* n. sp., and *Platyproteum* n. g. and comb. from North-Eastern Pacific peanut worms (Sipuncula). J. Eukaryot. Microbiol. *56*, 428–439.
10. Leander, B.S. (2006). Ultrastructure of the archigregarine *Selenidium vivax* (Apicomplexa) – A dynamic parasite of sipunculid worms (host: *Phascolosoma agassizii*). Mar. Biol. Res. *2*, 178–190.
11. Patten, R. (1936). Notes on a new Protozoon, *Piridium sociabile* n.gen., n.sp., from the foot of *Buccinum undatum*. Parasitology *28*, 502–516.
12. Burki, F., Kaplan, M., Tikhonenkov, D.V., Zlatogursky, V., Minh, B.Q., Radaykina, L.V., Smirnov, A., Mylnikov, A.P., and Keeling, P.J. (2016). Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proc. Biol. Sci. *283*, 20152802.
13. Irwin, N.A.T., Tikhonenkov, D.V., Hehenberger, E., Mylnikov, A.P., Burki, F., and Keeling, P.J. (2019). Phylogenomics supports the monophyly of the Cercozoa. Mol. Phylogenet. Evol. *130*, 416–423.
14. Kwong, W.K., Del Campo, J., Mathur, V., Vermeij, M.J.A., and Keeling, P.J. (2019). A widespread coral-infecting apicomplexan with chlorophyll biosynthesis genes. Nature *568*, 103–107.
15. Janouskovec, J., Horák, A., Oborník, M., Lukes, J., and Keeling, P.J. (2010). A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc. Natl. Acad. Sci. USA *107*, 10949–10954.
16. Huang, Y., He, L., Hu, J., He, P., He, J., Yu, L., Malobi, N., Zhou, Y., Shen, B., and Zhao, J. (2015). Characterization and annotation of *Babesia orientalis* apicoplast genome. Parasit. Vectors *8*, 543.
17. Gardner, M.J., Bishop, R., Shah, T., de Villiers, E.P., Carlton, J.M., Hall, N., Ren, Q., Paulsen, I.T., Pain, A., Berriman, M., et al. (2005). Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. Science *309*, 134–137.
18. Brayton, K.A., Lau, A.O.T., Herndon, D.R., Hannick, L., Kappmeyer, L.S., Berens, S.J., Bidwell, S.L., Brown, W.C., Crabtree, J., Fadrosh, D., et al. (2007). Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. PLoS Pathog. *3*, 1401–1413.
19. Abrahamsen, M.S., Templeton, T.J., Enomoto, S., Abrahante, J.E., Zhu, G., Lancto, C.A., Deng, M., Liu, C., Widmer, G., Tzipori, S., et al. (2004). Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. Science *304*, 441–445.
20. Toso, M.A., and Omoto, C.K. (2007). Gregarina niphandrodes may lack both a plastid genome and organelle. J. Eukaryot. Microbiol. *54*, 66–72.
21. Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D.M., Breiner, H.-W., and Richards, T.A. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol. Ecol. *19* (*Suppl 1*), 21–31.
22. Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.
23. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.
24. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.
25. Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.
26. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. *8*, 1494–1512.
27. Chevreux, B., Wetter, T., and Suhai, S. (1999). Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proc. 99th German Conference on Bioinformatics *99*, 45–56.
28. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.
29. Roure, B., Rodriguez-Ezpeleta, N., and Philippe, H. (2007). SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol. Biol. *7* (*Suppl 1*), S2.
30. Laetsch, D.R., and Blaxter, M.L. (2017). BlobTools: interrogation of genome assemblies. https://f1000research.com/articles/6-1287.
31. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212.

32. Andrews, S. (2018). FastQC. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

33. Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614–620.

34. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

35. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

36. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

37. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

38. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

39. Rambaut, A. (2007). FigTree, a graphical viewer of phylogenetic trees. https://beast.community/figtree.

40. Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics *25*, 2286–2288.

41. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

42. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE *5*, e9490.

43. Whelan, S., Irisarri, I., and Burki, F. (2018). PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. Bioinformatics *34*, 3929–3930.

44. Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. *340*, 783–795.

45. Sonnhammer, E.L., von Heijne, G., and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. *6*, 175–182.

46. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. *35*, W182–W185.

47. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

48. Csurös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics *26*, 1910–1912.

49. Aurrecoechea, C., Barreto, A., Basenko, E.Y., Brestelli, J., Brunk, B.P., Cade, S., Crouch, K., Doherty, R., Falke, D., Fischer, S., et al. (2017). EuPathDB: the eukaryotic pathogen genomics database resource. Nucleic Acids Res. *45* (D1), D581–D591.

50. Wang, H.-C., Minh, B.Q., Susko, E., and Roger, A.J. (2018). Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst. Biol. *67*, 216–235.

51. Ševčíková, T., Horák, A., Klimeš, V., Zbránková, V., Demir-Hilton, E., Sudek, S., Jenkins, J., Schmutz, J., Přibyl, P., Fousek, J., et al. (2015). Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? Sci. Rep. *5*, 10134.

52. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. *35*, 518–522.

53. Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29–W37.

54. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589.

55. Parsons, M., Karnataki, A., Feagin, J.E., and DeRocher, A. (2007). Protein trafficking to the apicoplast: deciphering the apicomplexan solution to secondary endosymbiosis. Eukaryot. Cell *6*, 1081–1088.

56. Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. *16*, 157.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Triton X-100	Sigma-Aldrich	cat. no. T9284
dNTP mix	Fermentas	cat. no. R0192
First-strand buffer	Invitrogen	cat. no. 18064-014
EB solution	QIAGEN	cat. no. 19086
Superscript II reverse transcriptase	Invitrogen	cat. no. 18064-014
Recombinant Ribonuclease Inhibitor	Invitrogen	cat. no. 10777019
Betaine	Sigma-Aldrich	cat. no. 61962
Magnesium chloride	Sigma-Aldrich	cat. no. M8266
KAPA HiFi HotStart ReadyMix (2 ×)	KAPA Biosystems	cat. no. KK2601
Ampure XP beads	Beckman Coulter	cat. no. A 63881
UltraPure DNase/RNase-Free Distilled Water	Thermofisher	cat. no. 10977023
Phusion High-Fidelity PCR Master Mix with HF Buffer	NEB	cat. no. M0531S
Ethanol 99.5% (vol/vol)	Kemethyl	cat. no. SN366915-06
DTT	Invitrogen	cat. no. 18064-014
Critical Commercial Assays		
AllPrep DNA/RNA Mini Kit	QIAGEN	Cat. No. 80204
Nextera XT	Illumina	FC-131-1024
Nextera	Illumina	FC-131-1024
Deposited Data		
Raw sequencing reads	This paper	NCBI SRA PRJNA5399860
Piridium plastid genome	This paper	GenBank MK962129
Experimental Models: Organisms/Strains		
Common whelk (*Buccinum undatum)*	This paper	N/A
Earthworm (*Lumbricus terrestris)*	This paper	N/A
Peanut worm (*Sipuncula* sp)	This paper	N/A
Skeleton shrimp (*Caprella californica)*	This paper	N/A
Polychaete worm (*Platyneries bicanaliculata)*	This paper	N/A
Bamboo worm (*Axiothella rubrocincta)*	This paper	N/A
Piridium sociabile	This paper	N/A
Platyproteum sp	This paper	N/A
Lecudina tuzetae	This paper	N/A
Monocystis agilis	This paper	N/A
Pterospora schizosoma	This paper	N/A
Heliospora capraellae	This paper	N/A
Oligonucleotides		
TAReuk454FWD1 (5′-CCAGCA(G⁄C)C(C⁄T) GCGG- TAATTCC-3′)	[21]	N/A
TAReukREV3 (5′-ACTTTCGTTCTTGAT(C⁄T) (A⁄G)A-3′)	[21]	N/A
Oligo-dT30VN (5′–AAGCAGTGGTATCAAC GCAGAGTACT30VN-3′)	[22]	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
IS-PCR oligo (5′-AAGCAGTGGTATCAACG CAGAGT-3′)	[22]	N/A
TSO (5′-AAGCAGTGGTATCAACGCAGAGT ACATrGrG+G-3′)	[22]	N/A
Software and Algorithms		
Trimommatic	[23]	http://www.usadellab.org/cms/?page=trimmomatic
Trinity	[24]	https://github.com/trinityrnaseq/trinityrnaseq/wiki
PRINseq	[25]	http://prinseq.sourceforge.net/index.html
Transdecoder	[26]	https://github.com/TransDecoder/TransDecoder/wiki
MIRA4	[27]	https://sourceforge.net/p/mira-assembler/wiki/Home/
Bowtie2	[28]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Geneious	https://www.geneious.com	N/A
SCaFoS	[29]	http://megasun.bch.umontreal.ca/Software/scafos/scafos.html
BlobTools	[30]	https://blobtools.readme.io/docs
BUSCO	[31]	https://busco.ezlab.org
PEAR	[32]	https://cme.h-its.org/exelixis/web/software/pear/
FastQC	[33]	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
BLAST	[34]	https://blast.ncbi.nlm.nih.gov/
MAFFT	[35]	https://mafft.cbrc.jp/alignment/software/
trimAl	[36]	http://trimal.cgenomics.org/
IQ-Tree	[37]	http://www.iqtree.org/
RAxML	[38]	https://sco.h-its.org/exelixis/web/software/raxml/
FigTree	[39]	https://beast.community/figtree
Phylobayes	[40]	http://www.atgc-montpellier.fr/phylobayes/
CD-HIT	[41]	http://weizhongli-lab.org/cd-hit/
FastTree	[42]	http://www.microbesonline.org/fasttree/
PREQUAL	[43]	https://github.com/simonwhelan/prequal
Divvier	https://github.com/simonwhelan/Divvier	N/A
SignalP	[44]	http://www.cbs.dtu.dk/services/SignalP/
TMHMM	[45]	http://www.cbs.dtu.dk/services/TMHMM/
Pfam	http://pfam.xfam.org/search/sequence	https://pfam.xfam.org/
Kyoto Encyclopedia of Genes and Genomes (KEGG)	[46]	https://www.genome.jp/kegg/
Diamond	[47]	https://github.com/bbuchfink/diamond
Count	[48]	http://www.iro.umontreal.ca/~csuros/gene_content/count.html

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Varsha Mathur (varsha.mathur@botany.ubc.ca). This study did not generate new unique reagents.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Monocystis agilis was isolated from the seminiferous vesicles of earthworms (*Lumbricus terrestris*), purchased from *Berry's Bait and Tackle,* Richmond, British Columbia, Canada in November 2017. *Platyproteum* sp. was isolated from the gut of a peanut worm (Sipuncula) that was collected from Sint Joris Bay, Curaçao in April 2018. *Lecudina tuzetae* and *Heliospora caprellae* were isolated from the guts of the animals *Platyneries bicanaliculata,* and *Caprella californica* respectively. They were both collected at low tide from Calvert Island, BC, Canada in June 2018. *Pterospora schizosoma* was isolated from the gut of a bamboo worm, *Axiothella rubrocincta,* that was collected from Friday Harbour, Washington, USA in June 2018. All of these parasites were collected in the trophozoite life stage (large feeding cells). Trophozoites were released into autoclaved filtered seawater by teasing apart the intestines/seminal vesicles of the respective hosts with pointed forceps.

Piridium sociabile was isolated from the common whelk, *Buccinum undatum*, that was collected using dredges across Breidafjördur, west coast of Iceland (65° 7.576'N; 22° 44.738'W). Whelks were sedated using 0.1% $MgSO_4$ in seawater for 1-2 hours, then examined for the presence of *Piridium* cysts on the surface of the foot using a dissection microscope. Mature (large) cysts were gently squeezed with pointed forceps until the *Piridium* gamonts were released. The resulting exudate was collected into concave glass spot plates containing filtered seawater and rinsed with autoclaved seawater three times to remove host tissues and mucous.

METHOD DETAILS

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 [22]. RNA and DNA sequencing libraries were both prepared using Illumina Nextera XT and Nextera protocol respectively, and sequenced using 2 × 300bp Illumina MiSeq (DNA) and 2x100bp Illumina HiSeq 2000 run (RNA). Both RNA and DNA reads were adaptor and quality trimmed with Trimommatic [23]. RNA reads were further processed to remove low complexity regions using PRINTseq [25] and were assembled into transcripts using Trinity v2.4 (with default settings) and translated into protein sequences using Transdecoder v.5 [24, 26].

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 [27]. This assembly was validated by mapping of the reads back to the assembly by Bowtie2 [28]. 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 40 × objective and a Sony α6000 camera. Single trophozoite cells were picked using a glass capillary micropipettes and transferred to a 0.2 mL 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 [22]. The cDNA concentration was quantified on a Qubit 2.0 Fluorometer (Thermo Fisher Scientific Inc.).

Prior to high-throughput sequencing, 1μL 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 [21]. 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 [34]. 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 [32, 33]. The adaptor and primer sequences were trimmed using Trimmomatic v0.36 and the transcriptomes were assembled with Trinity v2.4.0 [23, 24]. 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 [30, 34]. Coding sequences were predicted using a combination of TransDecoder v3.0.1 and similarity searches against the Swiss-Prot database [26]. Assessment of the quality of the assembly and annotation of the transcriptomes (including *Piridium*) was carried out using BUSCO [31].

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* [49]. All transcriptomes were comprehensively searched for a set of 263 genes that have been used in previous phylogenomic analyses [12, 13]. 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 [34]. The hits were then filtered using an e-value threshold of 1e-20 and a query coverage of 50%. Each of the genesets was then aligned using MAFFT L-INS-i v7.222 and trimmed using trimAl v1.2 with a gap-threshold of 80% [35, 36]. 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 [37, 38]. The resulting trees were manually scanned in FigTree v1.4.2 and contaminants and paralogous sequences were identified and removed [39]. 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 [29]. 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+Γ4+F model as implemented in IQ-TREE (model LG+Γ4+F yielded identical topology) [38]. Statistical support was inferred using 500 bootstrap replicates using the LG+C40+Γ4+F PMSF profiles, and 1000 bootstrap replicates using the LG+Γ4+F model in RAxML [37, 50]. The Bayesian tree was computed using Phylobayes [40] under the GTR-CAT model with constant sites removed from the analyses. Four independent chain were run for 9 thousand generations and converged with maxdiff = 0.19 (20% burning) (Figure 1G).

For the plastid based phylogenomic analyses, a previously published dataset [51] 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+Γ4+F substitution model with 500 bootstrap replicates [37] (Figure S2).

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, chrompodellids, 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) [34]. The database was subjected to CD-HIT with a similarity threshold of 85% to reduce redundant sequences and paralogs [41]. 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% [35, 36]. Maximum likelihood tree reconstructions were then performed with FastTree v. 2.1.7 using the default options [42]. 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 [36, 43]. 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%) [35]. 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 [38, 52].

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 [53]. 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 [38, 52]. 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 [39]. The remaining sequences were realigned and used to generate maximum likelihood phylogenies in IQ-TREE v.1.6.9 [38]. 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 [34, 54].

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 Nterminal 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 [44]. To predict putative Nterminal transmembrane domains, TMHMM v. 2.0 was used only on the first 100 amino acid residues of the transcript to improve prediction accuracy [45]. 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. [55] 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) [46]. 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 S2).

Ortholog identification and search for apicomplexan invasion/extracellular proteins

Orthofinder was used to infer orthologs within the apicomplexans, chrompodellids and *Platyproteum*, while *Oxytricha*, *Tetrahymena*, *Symbiodinium* and *Perkinsus* were used as a outgroup for the analyses [56]. Diamond searches were used to identify homologs between pairs of taxa [47]. The Orthofinder results were then analyzed using Dollo parsimony as implemented in the program Count to obtain estimates of gene gain and loss across different species [48]. Previously published sets of invasion (98 proteins from *Plasmodium falciparum* 3D7) and extracellular (722 proteins from diverse apicomplexans) proteins, as well as flagellar proteins, were then identified within orthogroups and their orthologs in the studied taxa were recorded (Table S3) [3, 4].

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical support for the phylogenomics tree was inferred using 500 bootstrap replicates using the LG+C40+Γ4+F PMSF profiles, and 1000 bootstrap replicates using the LG+Γ4+F model in RAxML [37, 50] (Figure 1). For the plastid based phylogenomic analyses, 500 non-parametric bootstrap replicates were used to assess support for the tree topology in RAxML (Figure S2) [37]. For the HMM search based phylogenies statistical support was assessed using 1000 ultrafast bootstrap pseudoreplicates in IQ-TREE [34].

DATA AND CODE AVAILABILITY

The genome and transcriptome sequencing reads are available in the NCBI Short Read Archive (SRA) NCBI: PRJNA539986. The *Piridium sociabile* plastid genome is available on GenBank: MK962129.