BMC Genomics

RESEARCH Open Access

Genome sequence of *Ophryocystis elektroscirrha*, an apicomplexan parasite of monarch butterflies: cryptic diversity and response to host-sequestered plant chemicals

Andrew J. Mongue[1*], Simon H. Martin[2], Rachel E. V. Manweiler[3], Helena Scullion[1], Jordyn L. Koehn[3], Jacobus C. de Roode[4] and James R. Walters[3]

Abstract

Apicomplexa are ancient and diverse organisms which have been poorly characterized by modern genomics. To better understand the evolution and diversity of these single-celled eukaryotes, we sequenced the genome of *Ophryocystis elektroscirrha*, a parasite of monarch butterflies, *Danaus plexippus*. We contextualize our newly generated resources within apicomplexan genomics before answering longstanding questions specific to this host-parasite system. To start, the genome is miniscule, totaling only 9 million bases and containing fewer than 3,000 genes, half the gene content of two other sequenced invertebrate-infecting apicomplexans, *Porospora gigantea* and *Gregarina niphandrodes*. We found that *O. elektroscirrha* shares different orthologs with each sequenced relative, suggesting the true set of universally conserved apicomplexan genes is very small indeed. Next, we show that sequencing data from other potential host butterflies can be used to diagnose infection status as well as to study diversity of parasite sequences. We recovered a similarly sized parasite genome from another butterfly, *Danaus chrysippus*, that was highly diverged from the *O. elektroscirrha* reference, possibly representing a distinct species. Using these two new genomes, we investigated potential evolutionary response by parasites to toxic phytochemicals their hosts ingest and sequester. Monarch butterflies are well-known to tolerate toxic cardenolides thanks to changes in the sequence of their Type II ATPase sodium pumps. We show that *Ophryocystis* completely lacks Type II or Type 4 sodium pumps, and related proteins PMCA calcium pumps show extreme sequence divergence compared to other Apicomplexa, demonstrating new avenues of research opened by genome sequencing of non-model Apicomplexa.

Keywords Monarch butterfly parasite, OE, Protist genomics, ATPase, Sodium potassium pump, Apicomplexan genomics, Cryptic species

*Correspondence:
Andrew J. Mongue
andrew.mongue@ufl.edu
[1]Department of Entomology and Nematology, University of Florida, Gainesville, USA
[2]Institute of Ecology and Evolution, University of Edinburgh, Edinburgh, UK
[3]Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, USA
[4]Biology Department, Emory University, Atlanta, USA

Introduction

Eukaryotic genomics has overwhelmingly focused on multicellular organisms [1], in spite of the staggering diversity of unicellular eukaryotes. When unicellular eukaryotes are studied, it is still mostly in connection to human health or economic interests. One of the most prominent examples of this limited and biased sampling is the Apicomplexa in the Kingdom Protista [2], a phylum of single-celled parasites including the organisms responsible for human diseases such as malaria, caused by *Plasmodium spp.*, and toxoplasmosis, from infections with *Toxoplasma gondii* [3]. Genomic studies of Apicomplexa have sequenced genomes that range from roughly 130 million bases on the high end [4] to a mere 9 million bases [5]; in other words, all Apicomplexa have small genomes compared to metazoans, but their genome sizes vary by an order of magnitude at least. Similarly, the gene content in these genomes is highly variable, from roughly 9,000 genes in *T. gondii* [6] to only 4,000 in the bovine parasite *Theileria orientalis* [7], a more than two-fold difference in gene content. However, with many unknowns about the genomics of apicomplexans that parasitize other vertebrates (see Levine 1986 for numerous examples) or invertebrates [9, 10], the overall patterns of genome size, gene content, and conservation of these features are open questions.

Ophryocystis elektroscirrha: a model non-model Apicomplexan parasite with open genetic questions

Among the invertebrate-infecting Apicomplexa, one of the best-studied is the neogregarine *Ophryocystis elektroscirrha* (Neogregarinorida: Ophryocystidae, McLaughlin and Myers 1970). It parasitizes butterfly species beginning when caterpillars ingest oocysts shed onto eggs and plant matter by infected adults. Within the host's gut, it reproduces first asexually then sexually as the host matures before forming dormant oocysts on the cuticle of the developing butterfly [9, 11]. Infected adults have shortened lifespans and decreased flight performance; those with the heaviest infections often fail to emerge from the pupa and quickly die [12, 13].

Ophryocystis elektroscirrha ostensibly infects a number of milkweed-feeding butterflies (Nymphalidae: Danainae). The initial description of this parasite listed both *Danaus plexippus*, the monarch butterfly, and *Danaus gilippus* as hosts [9]. Since then, similar apicomplexan infections have been reported in other members of the genus, including *D. eresimus, D. petilia* [14], and *D. chrysippus* [15]. Even apicomplexan infections of the more distantly related butterfly *Parthenos sylvia* and of moths in the genus *Helicoverpa* have been attributed to *O. elektroscirrha*-like parasites[16]. These diagnoses are based mainly on the gross morphological similarity of oocysts and may be limited by a paucity of informative morphological characters, however. As such, it is presently unclear whether *O. elektroscirrha* is a generalist lepidopteran parasite or if there is unrecognized parasite diversity between host species.

Another open question is how the parasite relates to the chemistry of the host's ecological niche. Monarch butterflies feed on several species of milkweed in the genus *Asclepias* which vary in the levels of phytochemicals they contain. Milkweeds produce cardiac glycosides called cardenolides, which are toxic to vertebrate predators [17] but sequestered by monarch caterpillars as a chemical defense [18]. Monarchs themselves remain largely unaffected thanks to a small set of mutations in their Na^+/K^+ – ATPase (sodium potassium pump) enzymes [19]. Intriguingly, parasite infection and virulence are strongly affected by plant chemistry; more toxic milkweeds, with more cardenolides, more strongly inhibit parasite growth [20]. Even indirect effects, such as the presence of secondary milkweed herbivores, have demonstrated effects on *O. elektroscirrha* infection dynamics in relation to plant chemistry [21].

It has already been demonstrated that other milkweed-feeding insects as well as their predators and metazoan parasites have evolved parallel amino acid changes in Na^+ /K^+ ATPases to tolerate these chemicals [22–24]. But it has been unclear if *O. elektroscirrha's* ATPases have evolved in response to milkweed biochemistry as well. Indeed, until recently Apicomplexa were thought to lack Na^+ ATPases, instead relying solely on Ca^{2+} ATPases, also known as calcium pumps [25]. However, it has been shown that this initial functional annotation (and likely comparative annotations relying on it) were incorrect, and many Apicomplexa do in fact possess distinct Na^+ ATPases [26, 27] that could be susceptible to cardiac glycosides.

To complement the wealth of existing ecological data available, and to begin to address the knowledge gaps highlighted above, we sequenced and annotated the genome of *Ophryocystis elektroscirrha*. We then set out to answer four questions: (1) How does the genome size and gene content of *O. elektroscirrha* compare to the few other invertebrate-infecting apicomplexan genomes available? (2) Can genome resequencing data of butterflies be used to diagnose their infection status? (3) Do the *Ophryocystis*-like infections in different *Danaus* butterfly species represent a single generalist parasite or separate parasite species? (4) How have ATPase genes evolved in *Ophryocystis* and Apicomplexa more generally, and is there evidence for adaptation to cardiac glycosides sequestered by their hosts? We find that *O. elektroscirrha* has a small genome, even by apicomplexan standards, which suggests high rates of gene loss across Apicomplexa. Our assembly provides a powerful tool for diagnosing butterfly infection status based on resequencing

data, as well as studying parasite diversity. With it, we show that the *Ophryocystis*-like parasites in two other *Danaus* species are significantly diverged from *O. elektroscirrha* and probably represent distinct species. Finally, we find that *Ophryocystis* may have adapted to host toxicity, but not in the manner that we predicted. We discuss these findings and highlight avenues for new research that our assembly opens up.

Methods

Parasite growth and propagation

We initially collected *O. elektroscirrha* (strain E41-1a) from a wild-caught eastern migratory monarch butterfly in October of 2017 in St. Marks, Florida, USA and propagated them in a laboratory setting; we fed second instar caterpillars a leaf disk containing a single oocyst to establish infections (following previously designed infection methods: [13, 28]). When the infected adult butterflies eclosed from their pupae, we froze them before collection of oocysts from the outsides of their bodies. Because oocysts are the result of meiotic cell division, this method results in a mix of related parasite genotypes rather than strictly identical clones but is ultimately the only way to generate enough parasite cells for DNA extraction and sequencing.

Concentration and purification of oocysts

We removed wings from the infected butterfly bodies and vortexed the bodies for 5 min in 100% ethanol in glass scintillation vials, a modification of de Roode et al. [28]. Oocysts, like the scales with which they associate, entered solution better in ethanol than water, generating a mix of both parasite oocysts and host scales. To separate scales from oocysts, we passed the solution through a 30 μm cell straining filter (Miltenyi Biotec, Bergisch Gladbach, Germany), which captured the scales (>100 μm length) while allowing the much smaller oocysts to pass through. We centrifuged the flow-through at 14,000 x g for 2 min to pellet the parasite oocysts and combined pellets across butterfly hosts to increase yield. Ultimately, we used oocysts from eight butterflies infected with the same initial parasite isolate (i.e. different oocysts from the same initial infected host) for DNA extraction.

Extraction and sequencing

The thick protein shell of the oocysts strongly inhibited lysing of parasite cells to access DNA. Thus, we needed to physically disrupt the oocysts prior to extraction. We ground the pellet with a Dounce homogenizer for 1.5 h in lysis buffer on ice, examining an aliquot under a light microscope every 15 min and stopped after nearly all of oocysts were visibly broken. Note that oocyst disruption began long before the 1-hour mark, so molecular protocols with lower input DNA requirements (e.g. PCR-based assays) would likely find success even with a greatly reduced disruption phase of this protocol. After this lengthy homogenization step, we followed the kit standard protocol for Omniprep extraction of genomic DNA (G-Biosciences, St. Louis, MO) and sequenced 250 basepair paired-end reads on an Illumina MiSeq with V3 chemistry.

To aid in assembly and annotation, we also generated RNA sequencing for *O. elektroscirrha* by extracting RNA from a heavily infected monarch pupa. We chose this host-stage for three reasons. First, *O. elektroscirrha* migrates to the cuticle and undergoes oocyst formation at this stage, guaranteeing that the parasite is transcriptionally active. Second, as this is the final active stage before going dormant in oocysts, it represents the peak number of parasite cells in the host. Finally, the level of infection is easily visible as dark spots on the green butterfly pupa [13], allowing us to select the most infected individual for extraction. Although contamination with host tissue and transcripts is practically unavoidable at this stage, we aimed to minimize contaminants by using dissecting scissors to target the dark aggregations of *O. elektroscirrha* while avoiding less infected host tissue. We extracted RNA using a Qiagen RNAeasy extraction kit (Hilden, Germany) and carried out Illumina 100 bp paired-end sequencing on a MiSeq with V3 chemistry.

Read processing and assembly

We were concerned that the tight association between host and parasite could still lead to accidental sequencing of host DNA in spite of our upstream attempts to separate the two tissue sources. As a final precaution against contamination, we first aligned raw sequenced reads to the monarch reference genome [29] with bowtie2 using the very-sensitive-local alignment algorithm [30] and only kept unmapped read pairs (--un-conc) for assembly. This methodology may weaken the power to detect horizontally transferred genes with conserved sequence but minimizes the chances of erroneously incorporating host sequence into the parasite assembly. With these high-confidence parasite reads, we sought first to set expectations prior to assembly using k-mer based methods. We counted k-mers using Jellyfish v2.2.6 with default parameters (k=21) [31] and then plotted the resultant k-mer frequency histogram and used this distribution to estimate genome size using custom scripts in R v3.3.3 [32]. Finally, we assembled these filtered reads with SPAdes v3.13 [33], using a k-mer coverage cutoff of 100 based on the characterization from the previous step. All other parameters were default settings for the tool.

Scaffolding

We aligned the RNAseq dataset to the newly generated assembly using TopHat v2.1.1 with a maximum intron

length of 500,000 bases [34] then used Rascaf v1.0.2 with default parameters [35] to scaffold the assembly based on RNA alignment. We assessed improvement to assembly summary statistics (N50 and contiguity) with QUAST v4.6.3 [36]. To assess the improvement in assembly of coding regions, we took both the original and scaffolded assembly through the annotation process described below, ending in evaluation with BUSCO v5.0 [37].

Annotation

We annotated the initial and RNA-scaffolded assemblies using the GenSAS web-based pipeline [38]. Prior to uploading, we trimmed the assemblies to remove contigs totaling fewer than 2,500 bases in length, both to meet requirements of the pipeline and because these contigs were unlikely to contain complete gene sequences. This approach resembles a recently successful effort to characterize the apicomplexan *Porospora gigantea* [5]. We then soft-masked the trimmed assembly after two independent rounds of repeat annotation, first with RepeatMasker v4.1.1 and then with RepeatModeler v2.0.1 before combining the two to mask bases for downstream analysis.

We employed two approaches to gene modeling, namely, BRAKER v2.1.1 [39] using alignment of RNA sequencing from the infected host pupa to the draft parasite genome, and GeneMark-ES v4.48 ab initio [40]. We evaluated both of these predicted gene sets using BUSCO v5.0 to search against the apicomplexa_odb10 database [37] and selected the gene set that maximized the number of identified single copy orthologs to use as the official gene set. Finally, to contextualize this new annotation with existing genome assemblies, we compared conserved orthologs against other gregarine Apicomplexa, namely the annotated protein sequences of *Gregarina niphandrodes* (unpublished, but available via the bioproject: PRJNA259233) and the two recently released *Porospora* genomes [41]. Again we used BUSCO v5.0's apicomplexa_odb10 database [37]. The aims of this analysis were twofold. First, we sought to evaluate how truly conserved the BUSCO dataset genes are, given that they were based on a different part of the apicomplexan tree. Second, we sought to understand patterns of gene conservation and turnover between distantly related species using ostensibly conserved genes.

Screening for infection in genome resequencing data

With the newly generated genome, we explored the potential to detect infection by *Ophryocystis* in genome resequencing data generated from butterfly tissues in order to compare host and parasite genetic variation. We used published Illumina resequencing datasets (100 or 150 bp paired-end) from seven species (Zhan et al. 2014; Martin et al. 2020, see supplemental files for sample details and accession numbers). We aligned Illumina reads to the masked *O. elektroscirrha* genome using bwa MEM version 0.7.17 [43] with default parameters. We used SAMtools version 1.9 [44] for conversion of SAM to BAM format, and Picard version 2.21.1 [45] SortSam and MarkDuplicates for sorting and removal of PCR duplicate reads. Given the massive overrepresentation of host butterfly DNA in the data, we took precautions to minimize misalignment of host DNA to the parasite genome. Specifically, only reads with a mapping quality of at least 60 were retained in the SAMtools step, and we subsequently removed all alignments of less than 100 bp in length, ignoring indels and soft clipping, using a custom Python script. We then computed mean read depth per scaffold using Mosdepth [46].

Screening for infection by detection of oocysts

For 18 *D. chrysippus* samples (Figure S2 and supplemental data tables) we had both Illumina resequencing data and preserved butterfly bodies, allowing us to compare the detection of infection using genomic data with the conventional approach of visually screening for oocysts under a microscope. Because the bodies were preserved in ethanol, we used a modified diagnostic procedure: A pipette was pressed against the abdomen and 10 µl of ethanol mixed with scales was pipetted onto a clean microscope slide, which was then viewed under 10x and 40x magnification for detection of oocysts. If too few scales were present on the slide, the procedure was repeated.

Identifying *Ophryocystis* scaffolds in a *Danaus chrysippus* assembly

Given that a *D. chrysippus* pupa used for a previous genome assembly [42] showed evidence for infection by *Ophryocystis* (see Results), we attempted to identify *Ophryocystis* scaffolds in the *D. chrysippus* assembly. We used a version of the assembly prior to removal of small scaffolds to ensure maximal recovery. We first generated a whole genome alignment between the *D. chrysippus* and *O. elektroscirrha* assemblies using minimap2 [47], with the 'asm20' parameter present, which is optimized for more dissimilar genomes. We then removed alignments less than 100 bp in length and those with sequence divergence (dv tag output by minimap2) > 0.15, based on visual inspection of the divergence distribution. Given the availability of a new highly complete *D. chrysippus* assembly from an uninfected individual [48], we generated a second whole genome alignment between the infected and uninfected *D. chrysippus* assemblies using the same procedure and filters. We reasoned that scaffolds representing *Ophryocystis* in the infected assembly should have strong homology to the *O. elektroscirrha* genome and little or no homology to the uninfected *D.*

chrysippus genome. However, given that repetitive sequences such as transposable elements (TEs) may be shared between host and parasite, we do not expect a complete lack of homology between parasite and host scaffolds. Based on visual inspection of the data, scaffolds were defined as confidently representing *Ophryocystis* if (1) alignments to *O. elektroscirrha* comprised at least half the scaffold length (after accounting for overlapping alignments), and (2) alignments to the uninfected *D. chrysippus* comprised less than a third of the scaffold length aligned to *O. elektroscirrha*.

As an additional line of evidence to exclude false positives, we considered read depth of Illumina reads from one infected and one uninfected adult butterfly (detected as described above). We reasoned that scaffolds representing *Ophryocystis* should have non-zero average read depth for reads from the infected adult, but zero read depth for reads from the uninfected adult. Given the low read depths (see Results) we required that scaffolds had a mean depth of at least 0.1 for the infected butterfly. Due to possible mis-mapping and shared repetitive sequences such as TEs, we relaxed the expectation that normalized read depth should be zero in the uninfected butterfly, and instead required that it must be lower than that in the infected butterfly.

Finally, we took these apicomplexan sequences through the same gene annotation pipeline as we used for our de novo *O. elektroscirrha* assembly. Ultimately, we compared the number and identity of BUSCO orthologs identified in each. Note however that this *D. chrysippus*-derived assembly is less contiguous owing to its incidental assembly within the host genome, so we discount cases of apparently missing orthologs as potential false negatives.

Phylogenetic diversity of ophryocystis

The ability to extract *Ophryocystis* sequence reads from genomic data of infected butterflies allowed us to investigate *Ophryocystis* diversity using previously published sequence data. We selected the twelve samples with the highest mean read depth for this purpose. Using the same filtered BAM files described above, we called haploid genotypes using bcftools v1.10.2 [44] using the mpileup, call, and filter tools to retain only non-indel genotypes with genotype quality (GQ) ≥ 20. Due to low sequencing depth (< 1x for most individuals) we could not reliably filter out potentially repetitive regions based on individual read depth. Instead, we excluded regions where the total read depth across all 12 individuals was > 30x, based on visual inspection of the genome-wide distribution. Our use of haploid genotypes assumes that each butterfly is infected by only a single parasite strain. Violation of this assumption would lead to a phylogenetic analysis in which each tip represents a combination of ancestries, which is in fact true for all genome-scale phylogenetics of recombining species. We therefore chose to use the distance-based neighbor-joining method, which is a genetic clustering algorithm that does not rely on an underlying model of a bifurcating tree. Pairwise genetic distances were computed using the distMat.py script (https://github.com/simonhmartin/genomics_general), considering only sites with genotypes for at least 9 of the 12 infected individuals. A neighbor-joining tree was generated using the BIONJ method implemented in splitsTree v4 [49].

Investigation of parasite ATPase evolution

We searched for signatures of co-evolutionary dynamics in the gene content of *Ophryocystis* in relation to its host's biochemical environment. To start, we obtained the protein sequence of the *Plasmodium falciparum* sodium potassium ATPase (accession AAF17245). This ATPase, a P-type ATPase 4, is often labeled as a Ca^{2+} ion pump, based on previous mis-annotation, but is likely an Na^{+} – ATPase as shown by more recent work in *Toxoplasma* [25–27]. We used BLAST+ (blastp -evalue 1e-10; Camacho et al. 2009) to extract homologous genes from the amino acid sequences of our newly generated *Ophryocystis* annotations as well those of *Gregarina niphandrodes* and *Porospora gigantea*-A. We then used a phylogenetic approach to place these unannotated gregarine sequences in the context of a recent and more comprehensive dataset of ATPases across Apicomplexa and Metazoa to infer ATPase type and function via sequence similarity [26]. We first aligned the gregarine protein sequences using MAFFT (as implemented in the Geneious software [51]), then manually aligned these sequences with the 265 amino acids corresponding to the highly conserved domains among ATPases analyzed by Lehane et al. [26], and finally trimmed the gregarine sequences to contain only those sites corresponding to the conserved 265 amino acids. Using this combined alignment of gregarine and other ATPases, we estimated a maximum likelihood phylogeny using the phangorn package in R, employing the LG+G(4)+I substitution model and 100 bootstrap replicates [52]. Subsequently, functional annotations were assigned to the gregarine sequences based on their placements within clades of distinct ATPases, as assessed by Lehane et al. Finally, we used phylogenetic relationships to infer functional classes of our uncharacterized gregarine ATPases in relation to established ATPases from other species.

Results

Raw sequence data composition

DNA sequencing produced 15,861,530 raw reads. Initial alignment showed 28% of total reads aligning to the host genome. We treated the former as butterfly sequences and the remaining 72% of unmapped reads (~ 11.4 million

reads) as putative parasite sequences. K-mer analyses with Jellyfish [31] suggested very little heterozygosity and a homozygous read depth of roughly 530x for these unmapped reads (Figure S1). Using this k-mer distribution and a custom R script, we calculated the expected genome size to be roughly 8.11 Mb, which would place *O. elektroscirrha* among the smallest sequenced apicomplexan genomes. See the Supplement for a more detailed explanation of calculations.

Assembly summary

An overview of the assembly is available in Table 1 (middle column). Initially, we assembled close to, but more than our k-mer based expectations: 8,864,135 bases. The assembly was contained in 909 contigs, with a minimum size of 128 basepairs and an N50 of 57,260 bases. We also noted that *O. elektroscirrha* is very GC poor (~27%), a feature characteristic of some other apicomplexan genomes. Before continuing with analyses, we needed to trim the smaller contigs from the assembly to meet criteria for annotation tools, so we set a minimum contig length threshold of 2,500 bases. This trimmed assembly totals 8,629,789 bases in 244 contigs, with an N50 of 58,409 bases. The loss of 234Kb of sequence (roughly 2.6% of the assembly) is a calculated tradeoff to format the genome for further analyses. Moreover, given the extremely high coverage and the method of extracting DNA from oocysts after sexual reproduction, it is likely that many of the very small contigs represent either read errors or regions of high genetic diversity that have enough coverage to not be discarded, but owing to their variation could not be collapsed into single sequences during assembly.

Finally, prior to annotation, we scaffolded the trimmed assembly using the RNA-seq data from an infected host pupa. We generated 31,574,399 read pairs, of which *Ophryocystis elektroscirrha* transcripts were a minority of sequences, as evidenced by an alignment rate of only roughly 7% of sequenced reads. Still, these ~2.3 million reads proved sufficient to improve the assembly. After scaffolding with Rascaf, the final assembly consists of 156 scaffolds with an N50 of 89,860 bases. This RNA scaffolding also improved the annotation of genes, as evidenced by an increase in identified BUSCOs reported below, by stitching together gene sequences split across contigs.

Repeat content

As expected of a very small genome, very little of the sequence was repetitive. In all, 1,014,979 bases (11.8% of the trimmed length), were masked. The majority of these, ~600 kb, were unclassified repeats, with the remaining ~400 kb being simple repeats.

Table 1 At-a-glance assembly and annotation statistics for the *Ophryocystis* sequences analyzed here. Raw size denotes the total length of bases *de novo* assembled or extracted from a host assembly in the case of the *O. elektroscirrha*-like assembly from *D. chrysippus*. We discarded sequences fewer than 2,500 bases in length prior to downstream gene annotation, yielding the final analyzed size of each genome, used for all subsequent analyses. We scaffolded the *O. elektroscirrha* assembly with RNA sequencing but did not perform comparable scaffolding for the *O. elektroscirrha*-like assembly due to divergence of sequences and lack of appropriate RNA data. We ran BUSCO on annotations to identify putative universally conserved single copy orthologs across Apicomplexa. Results follow the format: C – complete [S – single-copy, D – duplicated] F – fragmented, M – missing

	Ophryocystis elektroscirrha	*O. elektroscirrha*-like
Raw size	8,864,135 bp	8,799,632 bp
GC content	27.18%	26.93%
Analyzed size (final assembly)	8,631,036 bp	8,058,112 bp
Contigs	909	332
Contig N50	57,260 bp	49,155
Scaffolds	156	–
Scaffold N50	89,860 bp	–
Repeat content	11.76%	11.49%
Annotated genes (protein coding)	2,633 (2,591)	2,369 (2,280)
BUSCO apicomplexadb10 (n = 446)	C:80.9% [S:80.5%,D:0.4%], F:1.1%, M:18.0%	C:48.9% [S:48.0%,D:0.9%], F:11.2%, M:39.9%

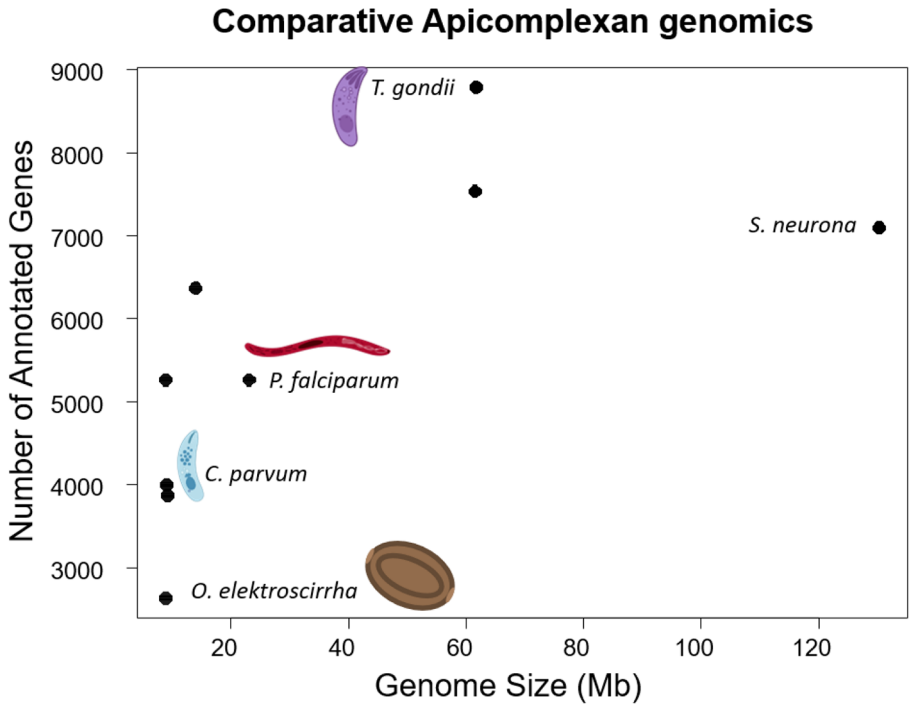

Fig. 1 A visual representation of genome size and gene content from Table 2, with notable taxa named and illustrated. Illustrations, except for *O. elektroscirrha* were created with BioRender.com. *Ophryocystis elektroscirrha* has the fewest annotated genes and one of the smallest overall genomes yet sequenced in Apicomplexa

Gene content

We employed two approaches to gene annotation, first using BRAKER to incorporate information from the alignment of RNAseq from an infected host to the *O. elektroscirrha* genome before RNA-scaffolding. With this method, we annotated 2,915 genes (encoding 3,122 proteins). This gene set contained 72.2% of the apicomplexa_odb10 BUSCO genes in a complete state, with a further 6% identifiable as fragmented. We also carried out ab initio annotation using GeneMark-ES without evidence beyond the genome sequence itself. We annotated 2,695 genes, 2,632 of which encoded a protein. This gene set contained far more BUSCOs, with 79.1% as complete genes and another 2% fragmented. As this method gave much better BUSCO results than BRAKER, we used it as the standard for subsequent annotations.

First, we reannotated the genome after scaffolding with RNAseq. In this assembly, GeneMark annotated 2,633 genes (2,591 protein coding). Note that although the total number of genes and the number of protein coding genes both decreased, the difference between the two decreased as well, as expected if RNA-scaffolding stitched together previously fragmented genes into complete coding sequences. This gene set contained 80.9% complete and 1.1% fragmented BUSCOs.

To contextualize this new assembly relative to previously studied Apicomplexa, we present summaries of genome size and gene content in both tabular (Table 2) and graphical form (Fig. 1). Additionally, we compared conserved gene content between *O. elektroscirrha*, *P. gigantea*-A, and *G. niphandrodes* (Fig. 2). In total, 309 of the 446 (69%) of BUSCO orthologs were conserved in all 3 genomes, with an additional 80 (18%) identifiable in two of three species. A further 32 orthologs (7%) were found only in one species, leaving 25 (6%) absent from all three. Despite *O. elektroscirrha* possessing roughly half the number of genes as the other two gregarines, its gene set is not a simple subset of either. Taken together this suggests that the core set of conserved apicomplexan orthologs is smaller than currently recognized and different lineages show unique patterns of gene loss and retention. Functional descriptions of BUSCOs unique to each lineage can be found in Table S1.

Genome resequencing data reveals infections in multiple *Danaus* species

The new OE genome allows us to detect parasite DNA in short-read genomic datasets, providing a route to screen for infection from sequenced samples without needing direct access to the butterfly itself. By analysing the depth of aligned Illumina reads (with stringent filtering, see Methods) from 38 wild-collected samples representing seven *Danaus* species, we found clear evidence for infection in five out of eight *D. plexippus* from Florida and

Table 2 *Ophryocystis elektroscirrha* in the context of other Apicomplexa. Genome size, annotated gene count, and gene density for a selection of published apicomplexan species. *Ophryocystis elektroscirrha* has a smaller genome than most other species and contains the fewest protein-coding genes yet-described for an Apicomplexan. We have excluded the *O. elektroscirrha*-like assembly from this comparison, as our methods of identification and filtering for annotation make it more likely to be an incomplete sequence and annotation

Species	Order	Genome size (Mb)	Gene content	Gene density (genes/Mb)	Reference
Sarcocystis neurona	Eucoccidiorida	130.2	7,093	54.5	Blazejewski et al., 2015
Toxoplasma gondii	Eucoccidiorida	61.6	8,789	142.7	Yucesan et al., 2021
Neospora caninum	Eucoccidiorida	61.5	7,540	122.6	Bernâ et al., 2021
Plasmodium falciparum	Haemospororida	22.9	5,268	245.8	Gardner et al., 2002
Gregarina niphandrodes	Eugregarinorida	14	6,375	455.4	Unpublished, bioproject: PRJNA259233
Cryptosporidium parvum	Eucoccidiorida	9.2	3,870	425.3	Abrahamsen et al., 2004
Theileria orientalis	Piroplasmida	9.0	4,002	444.7	Hayashida et al., 2012
Ophryocystis elektroscirrha	Neogregarinorida	8.8	2,633	299.2	This article
Porospora gigantea	Eugregarinorida	8.8	5,270	598.9	Boisard et al., 2022

Fig. 2 Conserved gene content overlap in sequenced gregarine Apicomplexa. Venn diagrams show the overlap in BUSCO orthologs identified from the apicomplexaodb10 dataset. Single, duplicated, and fragmented genes were all counted as present. Three distantly related gregarines all possess a large core of genes, 309 of 446 in the dataset. Moreover, *O. elektroscirrha* shares different sets of orthologs with each of the two species, suggesting independent lineage-specific gene loss across this group. Functional classifications of BUSCOs unique to each lineage can be found in the supplement (Table S1)

Ecuador, ten out of eighteen *D. chrysippus* from Kenya, and two out of two *D. petilia* from Australia (Figure S1, supplemental table). The remaining species were each represented by just one or two samples, so the absence of infections in these samples does not rule out that infections occur in the wild. The longest genome scaffolds provide the most robust evidence for infection. Shorter scaffolds show variable read depths in infected samples, and some have non-zero read depth in uninfected samples (Figure S1), likely indicating some shared repetitive DNA between the host and parasite genomes. Read depths tend to be low (<1X on average in most cases, compared to 10-25X coverage of host DNA), indicating that parasite DNA is far less abundant than host DNA, as expected of incidental sequencing without oocyst concentration or manual disruption.

Many of the samples considered were sequenced by other research groups, but the availability of bodies for 18 of the *D. chrysippus* samples allowed us to compare the accuracy of infection screening using genomic data versus the conventional method of microscopic detection of oocytes. This revealed nearly 100% correspondence, with nine individuals identified as infected using both methods, one with weak evidence for infection from genomic data but not from oocyst identification, and

eight identified as uninfected using both methods (Table S2). This implies that screening based on sequence data is at least as sensitive as the conventional approach.

A diverged *Ophryocystis* sequence from a related butterfly

Using the above *O. elektroscirrha* assembly, we scanned for apicomplexan scaffolds in a previous genome assembly of a *Danaus chrysippus* sample that had tested positive for infection (sample RFK001). In total, we extracted 822 sequences, totaling 8,799,632 bases (N50=44,501) (Figure S3). This would account for almost an entire genome if the sequences belong to *O. elektroscirrha*, but from the start, this conspecific status was dubious. Absolute divergence between sequences was 0.05. This substantial dissimilarity between sequences motivated further analyses.

To further investigate functional divergence, we used the same GeneMark approach as above to annotate the *Ophryocystis* sequence pulled from the *D. chrysippus* assembly. An overview of this assembly is shown on the right column of Table 1. We place less emphasis on the raw gene counts and number of BUSCOs missing from this assembly, as it is more fragmented than the *O. elektroscirrha* assembly and more of it had to be filtered out for analyses, resulting in only 8.06 Mb available for gene annotation. Nonetheless, we identified 2,369 genes (2,280 protein coding) with 48.9% complete BUSCOs and 11.2% fragmented. In raw numbers, 102 orthologs were missing from this *D. chrysippus-derived* genome compared to the *D. plexippus* parasite. More intriguingly, the *D. chrysippus* parasite annotation contained 4 BUSCO orthologs that were not found in *O. elektroscirrha*. We attempted to validate these results by BLASTing the putatively missing BUSCOs in the more complete, unscaffolded assembly and found strong hits (e-value$<10^{-20}$) for two of the four genes. Thus, while some of these differences appear to be false negatives arising from the assembly process, there may have been independent trajectories of gene loss and retention even between these two much more closely related parasites, in addition to the broad differences between sequenced gregarines. The two genes that remain missing are, by BUSCO ID, 32173at5974 – Ribosomal protein L37a and 16057at5794 – Eukaryotic translation initiation factor 3 subunit I.

To explore the relationship between parasite lineages from different host species, we used the *Ophryocystis*-like sequence data extracted from twelve infected butterflies to build a neighbor-joining tree. We restricted our analysis to sites at which high quality genotypes were present for at least nine of the twelve individuals. This filtered alignment included 12,976 variants across 365,363 aligned sites (4% of the genome, which is unsurprising given the low coverage and consequent high missingness). Sequences derived from *D. plexippus* butterflies (n=4) form one clade which includes the *O. elektroscirrha* reference sequence (Fig. 3). Sequences from *D. chrysippus* (n=6) form a distinct clade with substantial divergence from *O. elektroscirrha* as expected based on comparison between the genome assemblies. The *D. petilia* parasites (n=2) are most closely related to each other, and sister to the clade of sequences from *D. chrysippus*. Indeed, all parasites form monophyletic clades based on the host species from which they were sequenced, a pattern consistent with a host specificity of parasite lineages that could lead to co-evolution and speciation.

ATPase evolution

Using a *P. falciparum* ATPase to BLAST the annotated genes, we identified ATPases in both *Ophryocystis* assemblies as well as *P. gigantea* and *G. niphandrodes*, neither of which has known cardiac glycoside associations. *Ophryrocystis elektroscirrha* has three ATPases identifiable in the current annotation: g016110, g003020, and g013730. Similarly, three genes are found in the *O. elektroscirrha*-like annotation: g013280, g012340, and g002420, with the latter likely fragmented as a result of the method of assembly. The other gregarines, *G. niphandrodes* and *P. gigantea* have 3 and 6 putative ATPases respectively. It is notable that the *Porospora gigantea-A* annotation shows more ATPases than the other gregarines, suggesting potential gene duplication, but given the nature of how *Porospora* was assembled, it is possible these are pooled genes from two distinct lineages [5].

We placed all of these sequences in the context of more robustly annotated ATPases using a maximum-likelihood tree of conserved amino acid subsequences of ATPases. At a coarse level, gregarines fit within established apicomplexan patterns. None of our query ATPases belong to the ENA, or Type II ATPases, which have never been reported in other Apicomplexa. Moreover, all surveyed gregarines possess SERCA and PMCA calcium ATPases and at least some species possess ATP-4 sodium ATPases (Fig. 4). The exceptions to this general pattern are more interesting, and both involve *Ophryocystis*. First, neither *O. elektroscirrha* nor *O. elektroscirrha-like* possess ATP-4 sodium ATPases. And although both lineages have a pair of ATPases that fall within the PCMA family, the branch lengths compared to the rest of the phylogeny are very long. This is not merely a matter of overall sequence divergence from other taxa, as the SERCA ATPases of *Ophryocystis* are not exceptionally different from the rest of Apicomplexa. Thus, there appears to be a unique dynamic among the PMCA ATPases.

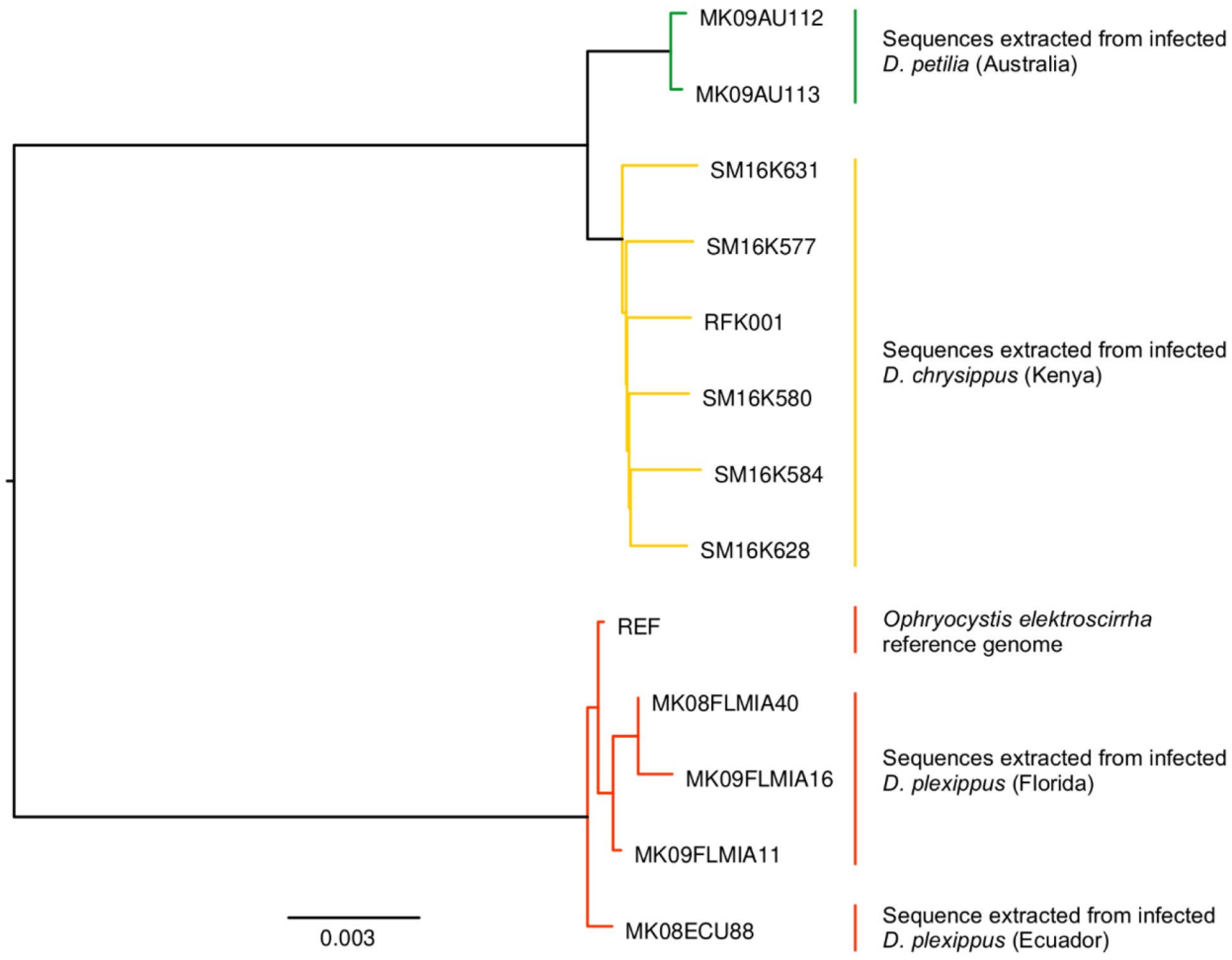

Fig. 3 A neighbor-joining tree of *Ophryocystis* sequences pulled from the sequencing of various milkweed butterflies, including both best-studied host *Danaus plexippus* and other related species. Branch lengths are proportional to sequence changes. Parasite sequences cluster perfectly with host species and *Ophryocystis* samples collected from *D. plexippus* form a distinct clade from those found in other *Danaus* species

Discussion

Here we report the assembly and gene annotation of the neogregarine parasite, *Ophryocystis elektroscirrha*. This species, along with other invertebrate pathogens, has largely been overlooked by modern genetic and genomic research. As such, the molecular characterization of this parasite has immediate value to both the broad understanding of apicomplexan biology and the specific host-parasite relationships between milkweed butterflies and *Ophryocystis.*

Ophryocystis elektroscirrha in relation to other Apicomplexa

Apicomplexa are an ancient and diverse clade of eukaryotes, about which very little is known outside of human-relevant pathogens [53]. As such, generating *a priori* expectations for genome size and content, as well as *a posteriori* assessment of an assembly and annotation, are difficult tasks. At the least, existing apicomplexan genomes give us some bounds for expectations. In comparison to other identified Apicomplexa, *Ophryocystis elektroscirrha* is the second smallest reported genome to-date at just under 9 megabases total, barely larger than the genomes of *Porospora* spp. [5]. Similarly, it has the fewest genes, but overall gene density (genes per megabase of sequence) is in the middle of the range of studied species.

Beyond these very coarse metrics though, finding similarities in genomic organization has generally proved difficult with Apicomplexa [54], likely for both biological and methodological reasons. On the biological side, parasites often have reduced genome sizes and gene counts compared to free-living organisms, owing to both increased selection for efficiency in replication and relaxed pressures to maintain molecular mechanisms that overlap with resources found in their hosts [55, 56].

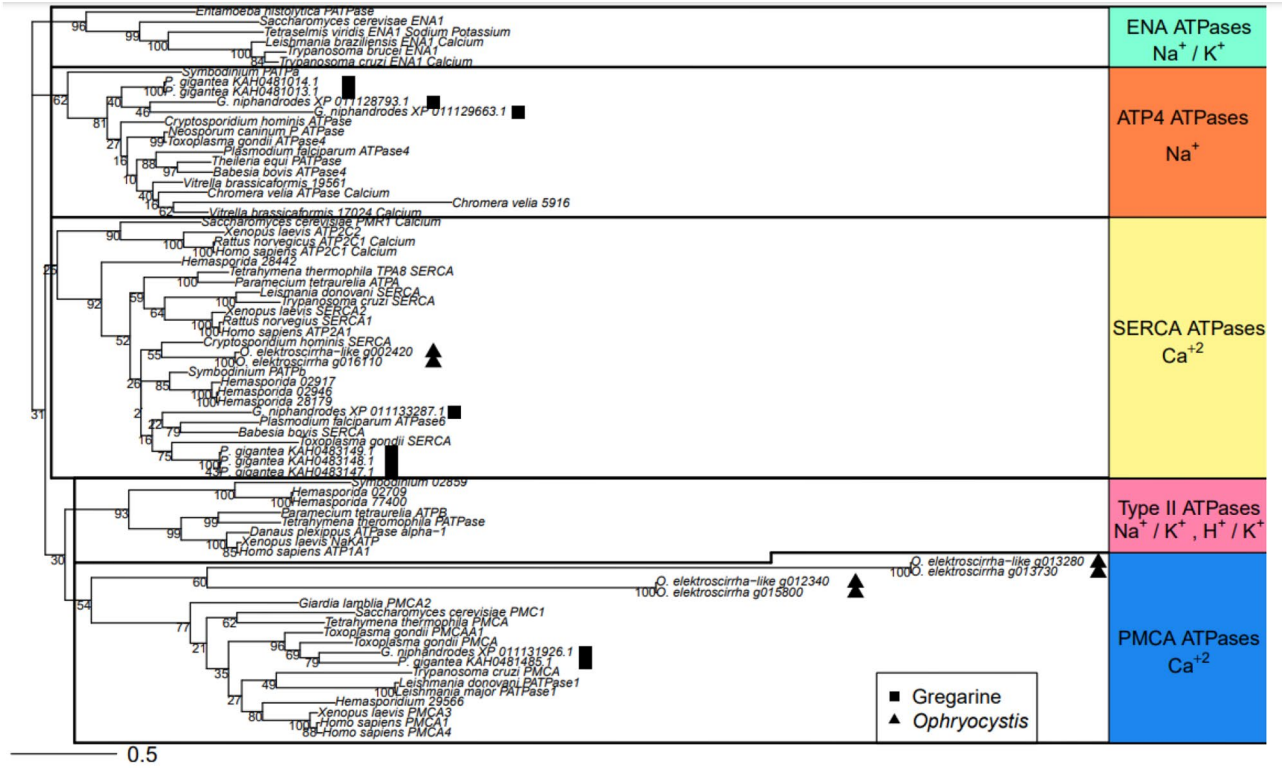

Fig. 4 Sequence-based categorization of gregarine ATPases. We used a set of highly conserved amino acid subsequences of ATPases, adding genes from *P. gigantea*, *G. niphandrodes*, *O. elektroscirrha*, *O. elektroscirrha-like* to a dataset from Lehane et al. [26] and generated a maximum likelihood tree with bootstrap support for nodes. As a group, the sequenced Gregarine ATPases (squares) do not differ from other studied Apicomplexa. None of these taxa possess Type II or ENA ATPases. At least some sampled gregarines have PMCA, SERCA, and ATP-4 type ATPases (squares), but *Ophryocystis* lineages (triangles) lack this last family. Additionally, *O. elektroscirrha* and *O. elektroscirrha-like* have two PMCA-like ATPases that show substantial sequence divergence from other members of the gene family. Branch lengths are proportional to sequence changes

Indeed, such explanations have long been invoked to explain apicomplexan genome sizes and gene contents [57]. Because these selective pressures are happening independently in different apicomplexan lineages, co-evolution with different hosts should result in different patterns of gene loss between parasite species. The end result is that attempting to find conserved sets of genes in the same order along chromosomes (i.e. synteny) between apicomplexan parasites has proved challenging [54].

On the methodological side, assessment of gene content is limited by available data. We recovered ~ 80% of expected Apicomplexan BUSCOs (i.e., "conserved" orthologs) in our annotated gene set for *O. elektroscirrha*, a low BUSCO score by most metazoan standards. From that perspective, such a high proportion of missing orthologs could be indicative of an incomplete or incompletely annotated assembly; however, the set of genes considered as "universal single copy" for a clade of organisms is defined only by existing genetic data [58]. Thus, genes that appear universally conserved within the small handful of well-characterized Apicomplexa may be truly lost in the *Ophryocystis* lineage. Indeed, a recent study of another gregarine parasite genus, *Porospora*, generated assemblies for two species, both of which with roughly ~ 70% of expected apicomplexan BUSCOs; a broader comparison in that same study identified 83% of the expected BUSCOs in *Gregarina niphandrodes* [5]. In that context, the conserved gene content of *O. elektroscirrha* fits well within gregarine expectations.

When we directly compared the overlap of BUSCOs identified in *Porospora* and *G. niphandrodes* to *O. elektroscirrha*, we found that only a small fraction (~ 6%) were truly absent from all three. More commonly, genes were conserved in only one or two species (26%). *Ophryocystis elektroscirrha*, despite having roughly half the annotated genes as the other two genera, was not merely a subset of either species in gene content. Instead, it displayed independent overlaps with the other two, as expected if co-evolution with different hosts and environments has driven unique patterns of gene loss in each lineage. More generally, these results suggest that the gregarine apicomplexans have a smaller set of conserved genes than currently recognized for the better-studied clades of Apicomplexa.

Cryptic diversity of milkweed butterfly parasites

Using the *O. elektroscirrha* genome, we were able to recover an apicomplexan genome from another milkweed butterfly, *Danaus chrysippus*. This sequence was highly diverged from the *O. elektroscirrha* genome, with absolute sequence divergence at roughly 5%. This number, which is comparable to the level of divergence between the two host butterfly species, may even be an underestimate, owing to our method of sequence discovery. Because we used similarity to the *O. elektroscirrha* genome to find new sequences, any regions that are exceptionally quickly evolving would not be detected via this method. Nevertheless, the comparable genome size suggests we are capturing the vast majority if not the entire parasite sequence.

We took these related *Ophryocystis* sequences through the same gene annotation pipeline we used for *O. elektroscirrha*. These new sequences were more fragmented, and thus unsurprisingly, we annotated fewer genes from the *D. chrysippus* derived parasite than from the *O. elektroscirrha* assembly. As such, apparent absences from the *chrysippus*-derived genome may be false negatives. In the other direction however, two apicomplexan BUSCOs were identified in the *chrysippus* parasite sequences that were absent from the *O. elektroscirrha* annotation. One of these is a translation initiation factor; the other is ribosomal protein L37. It is unclear why either of these should be lineage-specific and it is possible that these also represent false negatives, since both genomes were trimmed of short sequences prior to annotation. However, it is not outside the realm of possibility that these lineages may have different constraints on gene content if they consistently associate with different host species. In particular, other Apicomplexa have been observed to have incomplete or pseudogenized ribosomal components [59].

To explore the host and parasite relationship, we expanded sampling to sequenced reads from multiple infected *D. plexippus* and *D. chrysippus* and created a phylogenetic tree to examine their relatedness. We recovered a pattern of reciprocal monophyly for parasites based on host species. This pattern, along with the sequence and potential gene content divergence all suggest that different species of host may harbor distinct species or at least substantially differentiated lineages of *Ophryocystis*.

Work using only the 18 S ribosomal RNA sequence found that parasites isolated from *Danaus plexippus* and the distantly-related moth *Helicoverpa amigera* clustered in a similar pattern [16], but could not explore this pattern at larger scale without genomic data. More to the point, it is less surprising to see differentiation between parasites of hosts separated by ~ 110 million years of evolution [61] than to see such a pattern within a single host genus. It raises the possibility that other species of *Danaus* with a reported apicomplexan parasite are infected by distinct species of *Ophryocystis*.

Indeed, earlier experimental evidence has hinted at such a possibility. In cross-infection experiments exposing *D. plexippus* and *D. gilippus* hosts to *Ophryocystis* collected from either an intra- or interspecific source, the parasites were most successful infecting the same species of host from which they were first collected [14]. In other words, *Ophryocystis* exhibits significant host-specificity. What remains to be seen is how *Ophryocystis* lineages have evolved with *Danaus*. It may be that host and parasite share very similar speciation histories, as seen in other systems (e.g. birds and lice: Hughes et al. 2007). Alternatively, given that many species of milkweed butterfly share the same host plants in sympatric ranges, host switching may be driven more by an ecology of opportunity than phylogenetic history.

Milkweed, butterflies, parasites, and ATPases

The interactions between milkweed-feeding insects and their food source chemistry are well-studied. Milkweed toxicity derives in large part from a class of cardiac glycoside compounds that bind to and inhibit sodium potassium pump (Na^+/K^+ ATPase) proteins of the animals that ingest them [62]. For multicellular animals, sodium potassium pumps are key to the process of establishing an ion gradient across cell membranes and in particular allow for proper transmission of electrical signaling at the intercellular level. To maintain this crucial function, many milkweed-feeding insects have evolved a small set of amino acid substitutions in their ATPase sequence that confer resistance to cardiac glycoside binding. Unrelated herbivores such as butterflies and beetles show two convergent changes, a valine and histidine substitution in the α subunit of their Type II ATPase [19, 22]. Even more distantly related taxa, including a wasp parasitoid, a nematode parasite, and a bird predator of monarchs, all have similar substitutions in their ATPases [24]. Together these results suggest that a consistent selective pressure has driven a convergent molecular solution in independent lineages. What remains to be seen is if similar dynamics have occurred in non-metazoan members of the milkweed community.

Ophryocystis parasites of milkweed butterflies spend much of their life cycle in the larval gut or other host tissues [9] and are routinely exposed to cardiac glycosides. Experimental evidence shows that *O. elektroscirrha* growth is negatively impacted by the presence and concentration of cardiac glycosides [20] and that infected female *D. plexippus* (which would likely transmit *O. elektroscirrha* to their offspring) preferentially choose to lay eggs on milkweed with more cardiac glycosides when given a choice [63]. Thus, the hosts and their ingested

phytochemicals appear to exert a selective pressure on the parasite, but what their molecular targets are had yet to be explored.

Aside from a lack of sequence data, the main challenge to this line of research is that between Metazoa and Apicomplexa there is a lack of homology, in both protein sequence and function. First, although ATPases are an evolutionarily ancient class of proteins found in protists as well, they obviously cannot play roles in intercellular electrical signaling of a single-celled organism, so the mechanism of toxicity cannot be the same. But ATPases are still important regulators of cellular homeostasis with respect to salt and pH balance and have been suggested as targets for drug development [27, 64]. Even if their functions are different, inhibition of apicomplexan ATPases would still be detrimental to the organism.

In the above descriptions of ATPase – cardiac glycoside interactions in Metazoa, the target ATPase is always the Type II Na^+/K^+ ATPase. No Apicomplexa are known to possess this specific family of ATPase and, until recently, they were thought to lack any sort of Na^+ ATPase. More recently however, it has been shown in *Toxoplasma gondii* that the ATP-4 ATPase, previously thought to employ calcium, is in fact a sodium ATPase [26]. This family could be a candidate for the cardiac glycosides' target, given the conserved cation use. However, none of the ATPases had been characterized in gregarines.

We used conserved sequence domains across ATPases to characterize the specific families of ATPases present in *P. gigantea* and *G. niphandrodes*, *O. elektroscirrha*, and *O. elektroscirrha-like*. As a whole, the gregarines fit with better-studied Apicomplexa. None possess Type II or ENA ATPases, and all possess PMCA and SERCA calcium ATPases. Intriguingly though, both *Ophryocystis* lineages apparently lack ATP-4 sodium ATPases that are found in *Porospora* and *Gregarina*. As the only putative sodium ATPase in Apicomplexa, it is tempting to speculate that loss of this ATPase may have been related to cardiac glycoside presence in the host. A parasite ATPase that is routinely inhibited by the host's chemical environment is essentially non-functional and may be lost under mutation accumulation. Of course, this line of reasoning relies on only a small set of observations; it would be bolstered by discovering ATP-4 sodium ATPase genes in closer relatives to *Ophryocystis* that parasitize non-milkweed-feeding insects.

Considering the ATPases still present in *Ophryocystis*, we recovered three putative calcium pumps, a SERCA (localized to the endoplasmic reticulum within the cell) and two PMCA ATPases which are located on the plasma membrane of the cell. These should be considered as potential targets of cardiac glycosides, as some sodium pump blockers may have non-specific inhibitory action against calcium pump ATPases as well [65]. The PMCAs would be most likely to interact with cardiac glycosides, which canonically affect plasma membrane proteins [66]. In *Ophryocystis*, these proteins' sequences are very different from other sequenced Apicomplexa; in contrast, the SERCA ATPases of *Ophryocystis* do not show exceptional sequence divergence from orthologous genes. Thus, only the plasma membrane associated ATPases of *Ophryocystis* appear highly diverged. It will require more gregarine sequences for comparison, but these proteins are intriguing candidates for the target of cardiac glycoside toxicity in milkweed butterfly parasites.

Conclusions

Sequencing the genome of *Ophryocystis elektroscirrha* has yielded immediate insights into potential biochemical evolution driven by milkweed chemistry, cryptic variation in milkweed butterfly parasites, and the true extent of gene turnover across Apicomplexa. We present these results to facilitate further exploration of these parasites to give context to the large body of disease ecology research on this clade. We hope that our novel data collection methods, both in DNA extraction and parasite sequence screening, will aid in future work on poorly understood Apicomplexa.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09350-0.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

The authors wish to thank Adele Lehane, Adelaide Dennis, and David Heckel for insights into ATPase evolution; Sonia Altizer and Maria Luisa Muller Theissen for discussions about OE – butterfly interactions; and Jasmin Albert, Thomas Johnson, and Megan Hansen for additional help in protocol optimization of DNA extraction. Thanks to Isabelle Florent for additional context and insights on general apicomplexan genomics.

Authors' contributions

AJM oversaw project planning, wet lab work, assembly and annotation, comparative genomic analyses, and wrote the body of the manuscript. SHM conducted identification and phylogenetic analyses of the OE-like genome, developed and implemented read-based screening analyses, and contributed to the manuscript text .REVM conducted bioinformatic analyses scaffolding the primary assembly with RNA and evaluating assembly improvements. HS conducted phylogenetic comparisons and visual screening of parasite infection in butterfly samples. JLK optimized and carried out the DNA extraction protocol. JCdR provided parasite samples for sequencing and provided in-depth editing of the manuscript. JRW provided funding for sequencing, conducted ATPase protein evolution analyses and provided in-depth feedback. All authors reviewed the manuscript.

Funding

James Walters was supported by the US National Science Foundation grant NSF-ABI 1661454. Rachel Manweiler and Jordyn Koehn were supported by the Gould Summer Entomology Fellowship from the University of Kansas.

Data Availability
The genome assembly for *Ophryocystis elektroscirrha* can be found with the accession JAQIFP000000000. Raw DNA reads used to assemble the genome and RNAseq used to scaffold can be found with PRJNA906508. The annotation, along with the *O. elektroscirrha-like* assembly and annotation can also be found at https://doi.org/10.5281/zenodo.7817702; the assembly in particular we chose not to formally archive due to its potentially incomplete nature and uncertain taxonomic status. Accessions for the butterfly sequences used to screen for *Ophryocystis* reads can be found in the Supplement as well. Custom downstream analysis scripts are housed at https://github.com/amongue/OE_genome .

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors have no competing interests as defined by BMC.

Received: 10 February 2023 / Accepted: 29 April 2023
Published online: 24 May 2023

References

1. Hotaling Scott, Kelley Joanna L, Frandsen Paul B. Toward a genome sequence for every animal: Where are we now? Proceedings of the National Academy of Sciences. 2021;118:e2109019118.
2. Templeton TJ, Iyer LM, Anantharaman V, Enomoto S, Abrahante JE, Subramanian G, et al. Comparative analysis of apicomplexa and genomic diversity in eukaryotes. Genome Res. 2004;14:1686–95.
3. Levine ND. The Protozoan Phylum Apicomplexa: volume 2. CRC Press; 2018.
4. Blazejewski T, Nursimulu N, Pszenny V, Dangoudoubiyam S, Namasivayam S, Chiasson MA, et al. Systems-based analysis of the Sarcocystis neurona genome identifies pathways that contribute to a heteroxenous life cycle. MBio. 2015;6:e02445–14.
5. Boisard J, Duvernois-Berthet E, Duval L, Schrével J, Guillou L, Labat A, et al. Marine gregarine genomes reveal the breadth of apicomplexan diversity with a partially conserved glideosome machinery. BMC Genomics. 2022;23:1–22.
6. Yucesan B, Guldemir D, Babur C, Kilic S, Cakmak A. Whole-genome sequencing of a Toxoplasma gondii strain from a turkish isolate using next-generation sequencing technology. Acta Trop. 2021;218:105907.
7. Hayashida K, Hara Y, Abe T, Yamasaki C, Toyoda A, Kosuge T, et al. Comparative genome analysis of three eukaryotic parasites with differing abilities to transform leukocytes reveals key mediators of Theileria-induced leukocyte transformation. MBio. 2012;3:e00204–12.
8. Levine ND. The taxonomy of Sarcocystis (protozoa, apicomplexa) species. J Parasitol. 1986;:372–82.
9. McLaughlin RE, Myers J. Ophryocystis elektroscirrha sp. n., a Neogregarine Pathogen of the Monarch Butterfly Danaus plexippus (L.) and the Florida Queen Butterfly D. gilippus berenice Cramer1. J Protozoology. 1970;17:300–5.
10. Chen W. The life cycle of Ascogregarina taiwanensis (Apicomplexa: Lecudinidae). Parasitol Today. 1999;15:153–6.
11. Tanada Y, Kaya H. Insect Pathology Academic Press. Inct Harcaundt Brace Jovanavich Publisher San Diego; 1993.
12. Bradley CA, Altizer S. Parasites hinder monarch butterfly flight: implications for disease spread in migratory hosts. Ecol Lett. 2005;8:290–300.
13. de Roode JC, Yates AJ, Altizer S. Virulence-transmission trade-offs and population divergence in virulence in a naturally occurring butterfly parasite. Proc Natl Acad Sci. 2008;105:7489–94.
14. Barriga PA, Sternberg ED, Lefèvre T, de Roode JC, Altizer S. Occurrence and host specificity of a neogregarine protozoan in four milkweed butterfly hosts (Danaus spp). J Invertebr Pathol. 2016;140:75–82.
15. Ndatimana G, Kayitete L, Martin S, Smith DA, Hagenimana T, Nkundimana A et al. Morph frequencies, sex ratios and infections in Danaus chrysippus populations in Rwanda. Afr J Ecol. 2022;60:633-640
16. Gao K, Muijderman D, Nichols S, Heckel DG, Wang P, Zalucki MP, et al. Parasite-host specificity: a cross-infection study of the parasite Ophryocystis elektroscirrha. J Invertebr Pathol. 2020;170:107328.
17. Brower LP, Fink LS. A natural toxic defense system: cardenolides in butterflies versus birds. Ann N Y Acad Sci. 1985;443:171–88.
18. Parsons J. A digitalis-like toxin in the monarch butterfly, Danaus plexippus L. J Physiol. 1965;178:290.
19. Aardema ML, Zhen Y, Andolfatto P. The evolution of cardenolide-resistant forms of Na+, K+ - ATPase in Danainae butterflies. Mol Ecol. 2012;21:340–9.
20. de Roode JC, Pedersen AB, Hunter MD, Altizer S. Host plant species affects virulence in monarch butterfly parasites. J Anim Ecol. 2008;77:120–6.
21. De Roode JC, Rarick RM, Mongue AJ, Gerardo NM, Hunter MD. Aphids indirectly increase virulence and transmission potential of a monarch butterfly parasite by reducing defensive chemistry of a shared food plant. Ecol Lett. 2011;14:453–61.
22. Aardema ML, Andolfatto P. Phylogenetic incongruence and the evolutionary origins of cardenolide-resistant forms of Na+, K+ - ATPase in Danaus butterflies. Evolution. 2016;70:1913–21.
23. Pierce AA, de Roode JC, Tao L. Comparative genetics of Na+/K+-ATPase in monarch butterfly populations with varying host plant toxicity. Biol J Linn Soc. 2016;119:194–200.
24. Groen SC, Whiteman NK. Convergent evolution of cardiac-glycoside resistance in predators and parasites of milkweed herbivores. Curr Biol. 2021;31:R1465–6.
25. Krishna S, Woodrow C, Webb R, Penny J, Takeyasu K, Kimura M, et al. Expression and functional characterization of a Plasmodium falciparum Ca2+-ATPase (PfATP4) belonging to a subclass unique to apicomplexan organisms. J Biol Chem. 2001;276:10782–7.
26. Lehane AM, Dennis AS, Bray KO, Li D, Rajendran E, McCoy JM, et al. Characterization of the ATP4 ion pump in Toxoplasma gondii. J Biol Chem. 2019;294:5720–34.
27. Dick CF, Meyer-Fernandes JR, Vieyra A. The functioning of Na+-ATPases from Protozoan Parasites: are these pumps targets for antiparasitic drugs? Cells. 2020;9:2225.
28. de Roode JC, Gold LR, Altizer S. Virulence determinants in a natural butterfly-parasite system. Parasitology. 2007;134:657–68.
29. Gu L, Reilly PF, Lewis JJ, Reed RD, Andolfatto P, Walters JR. Dichotomy of Dosage Compensation along the neo Z chromosome of the Monarch Butterfly. Curr Biol. 2019;29:4071–4077e3.
30. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
31. Marcais G, Kingsford C, Jellyfish. A fast k-mer counter. Tutorialis e Manuais. 2012;1:1–8.
32. R Core Team. R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria URL http://wwwR-project.org/. 2017;:R Foundation for Statistical Computing.
33. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.
34. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:1–13.
35. Song L, Shankar DS, Florea L. Rascaf: improving genome assembly with RNA sequencing data. The plant genome. 2016;9:plantgenome2016–03.
36. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.
37. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38:4647–54.
38. Humann JL, Lee T, Ficklin S, Main D. Structural and functional annotation of eukaryotic genomes with GenSAS. Gene prediction.Springer; 2019. 29–51.
39. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2015;32:767–9.
40. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 2008;18:1979–90.

41. Zhan S, Zhang W, Niitepõld K, Hsu J, Haeger JF, Zalucki MP, et al. The genetics of monarch butterfly migration and warning colouration. Nature. 2014;514:317–21.
42. Martin SH, Singh KS, Gordon IJ, Omufwoko KS, Collins S, Warren IA, et al. Whole-chromosome hitchhiking driven by a male-killing endosymbiont. PLoS Biol. 2020;18:e3000610.
43. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26:589–95.
44. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
45. Wysoker A, Tibbetts K, Fennell T. Picard tools version 1.90. http://picard sourceforge net (Accessed 14 December 2016). 2013;107:308.
46. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018;34:867–8.
47. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
48. Singh KS, De-Kayne R, Omufwoko KS, Martins DJ, Bass C, Ffrench-Constant R et al. Genome assembly of Danaus chrysippus and comparison with the Monarch Danaus plexippus. G3. 2022;12:jkab449.
49. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006;23:254–67.
50. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:1–9.
51. Geneious Prime software, version 2022.2.2 (http://www.geneious.com)
52. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27:592–3.
53. Morrison DA. Evolution of the Apicomplexa: where are we now? Trends Parasitol. 2009;25:375–82.
54. DeBarry JD, Kissinger JC. Jumbled genomes: missing Apicomplexan synteny. Mol Biol Evol. 2011;28:2855–71.
55. Keeling PJ, Slamovits CH. Causes and effects of nuclear genome reduction. Curr Opin Genet Dev. 2005;15:601–8.
56. Sundberg L-R, Pulkkinen K. Genome size evolution in macroparasites. Int J Parasitol. 2015;45:285–8.
57. Keeling PJ. Reduction and compaction in the genome of the apicomplexan parasite Cryptosporidium parvum. Dev Cell. 2004;6:614–6.
58. Jauhal AA, Newcomb RD. Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. Mol Ecol Resour. 2021;21:1416–21.
59. Mathur V, Kwong WK, Husnik F, Irwin NA, Kristmundsson Á, Gestal C, et al. Phylogenomics identifies a new major subgroup of apicomplexans, marosporida class nov., with extreme apicoplast genome reduction. Genome Biol Evol. 2021;13:evaa244.
60. Heikkila M, Kaila L, Mutanen M, Pena C, Wahlberg N. Cretaceous origin and repeated tertiary diversification of the redefined butterflies. Proceedings of the Royal Society B: Biological Sciences. 2012;279:1093–9.
61. Hughes J, Kennedy M, Johnson KP, Palma RL, Page RD. Multiple cophylogenetic analyses reveal frequent cospeciation between pelecaniform birds and Pectinopygus lice. Syst Biol. 2007;56:232–51.
62. Agrawal AA, Petschenka G, Bingham RA, Weber MG, Rasmann S. Toxic cardenolides: chemical ecology and coevolution of specialized plant–herbivore interactions. New Phytol. 2012;194:28–45.
63. Lefèvre T, Oliver L, Hunter MD, De Roode JC. Evidence for trans-generational medication in nature. Ecol Lett. 2010;13:1485–93.
64. Yamasaki M, Takada A, Yamato O, Maede Y. Inhibition of na, K-ATPase activity reduces Babesia gibsoni infection of canine erythrocytes with inherited high K, low na concentrations. J Parasitol. 2005;91:1287–92.
65. Kelly RA, O'Hara DS, Canessa ML, Mitch WE, Smith TW. Characterization of digitalis-like factors in human plasma. Interactions with NaK-ATPase and cross-reactivity with cardiac glycoside-specific antibodies. J Biol Chem. 1985;260:11396–405.
66. Matsui H, Schwartz A. Mechanism of cardiac glycoside inhibition of the (Na+-K+)-dependent ATPase from cardiac tissue. Biochim et Biophys Acta (BBA)-Enzymology. 1968;151:655–63.

Publisher's Note