Protist, Vol. 166, 659–676, December 2015
http://www.elsevier.de/protis
Published online date 30 September 2015

Protist

ORIGINAL PAPER

Phylogeny, Morphology, and Metabolic and Invasive Capabilities of Epicellular Fish Coccidium *Goussia janae*

Sunil Kumar Dogga[a], Pavla Bartošová-Sojková[b], Julius Lukeš[b,c,d], and Dominique Soldati-Favre[a,1]

[a]Department of Microbiology and Molecular Medicine, University of Geneva. CMU, 1 Rue Michel-Servet, CH-1211 Geneva 4, Switzerland
[b]Institute of Parasitology, Biology Centre, Branišovská 31, České Budějovice (Budweis), Czech Republic
[c]Faculty of Science, University of South Bohemia, Branišovská 1645/31A, České Budějovice (Budweis), Czech Republic
[d]Canadian Institute for Advanced Research, 180 Dundas St W, Toronto, ON M5G 1Z8, Canada

Submitted June 29, 2015; Accepted September 15, 2015
Monitoring Editor: Frank Seeber

To fill the knowledge gap on the biology of the fish coccidian *Goussia janae*, RNA extracted from exogenously sporulated oocysts was sequenced. Analysis by Trinity and Trinotate pipelines showed that 84.6% of assembled transcripts share the highest similarity with *Toxoplasma gondii* and *Neospora caninum*. Phylogenetic and interpretive analyses from RNA-seq data provide novel insight into the metabolic capabilities, composition of the invasive machinery and the phylogenetic relationships of this parasite of cold-blooded vertebrates with other coccidians. This allows re-evaluation of the phylogenetic position of *G. janae* and sheds light on the emergence of the highly successful obligatory intracellularity of apicomplexan parasites. *G. janae* possesses a partial glideosome and along with it, the metabolic capabilities and adaptions of *G. janae* might provide cues as to how apicomplexans adjusted to extra- or intra-cytoplasmic niches and also to become obligate intracellular parasites. Unlike the similarly localized epicellular *Cryptosporidium* spp., *G. janae* lacks the feeder organelle necessary for directly scavenging nutrients from the host. Transcriptome analysis indicates that *G. janae* possesses metabolic capabilities comparable to *T. gondii*. Additionally, this enteric coccidium might also access host cell nutrients given the presence of a recently identified gene encoding the molecular sieve at the parasitophorous vacuole membrane.

Key words: Apicomplexa; Coccidia; *Goussia janae*; phylogeny; ultrastructure; invasion; central carbon metabolism.

Introduction

Members of the phylum Apicomplexa share distinctive morphological traits, the presence of a

[1]Corresponding author; fax +41-22-3795702
e-mail dominique.soldati-favre@unige.ch (D. Soldati-Favre).

http://dx.doi.org/10.1016/j.protis.2015.09.003
1434-4610/

set of unique organelles at the apical end being their prominent unifying feature. This apical complex includes the secretory organelles called the rhoptries and micronemes, the apical polar ring, and the conoid. The rhoptries and micronemes are unique secretory organelles that are associated with motility, adhesion, invasion and establishment of a parasitophorous vacuole. The conoid is a cone-shaped structure believed to play a mechanical role in the invasion of host cells. The parasites possess flattened membranous vesicles called alveoli, closely apposed beneath the plasma membrane. The composite structure of the parasite membrane and the inner membrane complex (IMC) is associated with a number of cytoskeletal elements, including actin, myosin, microtubules, and a network of intermediate filament-like proteins. The IMC plays a structural role and is an important scaffold element during cytokinesis, in addition to being involved in motility and invasion of the host cell (Harding and Meissner 2014). Most apicomplexans also have another unique structure, a chloroplast-like organelle called the apicoplast, a derived non-photosynthetic plastid (van Dooren and Striepen 2013).

The genus *Goussia* Labbé, 1896 (Apicomplexa, Eimeriorina) accommodates piscine and amphibian coccidia with thin-walled oocysts which lack micropyle and contain four dizoic bivalved sporocysts. Although members of the genus *Goussia* are mostly encountered in the intestine, infections of other organs, such as kidney and spleen are quite common. Most species have homoxenous life cycles, but a more complex cycle involving fish as definitive hosts and tubificids and other invertebrates as vectors and/or paratenic hosts was experimentally proven in a few cases (Dyková and Lom 1981; Lom and Dyková 1992; Fournie and Overstreet 1983; Steinhagen and Körting 1990).

Goussia janae was described from the intestine of dace *Leuciscus leuciscus* and chub *Squalius cephalus*. It is a causative agent of heavy infections manifested by microvillar atrophy of epithelial cells and subsequent formation of multiple secondary mucosal folds (Lukeš and Dyková 1990). The parasite has a strict seasonal occurrence, with immature oocysts being shed in spring from its fish host in feces or massively in casts. Sporulation is exogenous and at 10 °C takes about 48 hours (Lukeš and Dyková 1990; Lukeš and Starý 1992). *G. janae* development occurs inside the intestinal epithelium, yet in a characteristic epicellular location (also known as extracytoplasmic). This means that merogonial, gamogonial and early sporogonial stages are confined to the microvillar region of the epithelial cells, where they are covered by very closely apposed enterocyte and parasitophorous vacuole membranes (PVMs) (Lukeš and Starý 1992; Molnár 1996). The more frequent "monopodial" subtype of the epicellular interaction means that the coccidium forms a single zone of attachment with the host cell cytoplasm, whereas the so-called "spider-like" subtype describes stages positioned above the microvillar region, connected to the host cell only through several narrow zones of attachment (Bartošová-Sojková et al. 2015; Lukeš and Starý 1992).

In the 18S rDNA-based phylogeny, some piscine coccidians (e.g. *G. janae*, *G. ameliae*, *G. pannonica, G. szekelyi*) form basal lineages of the species-rich eimeriorinid clade, or of its sarcocystid subclade, which encompass economically important parasites of both homeothermic and poikilothermic vertebrate hosts (Bartošová-Sojková et al. 2015; Jirků et al. 2002; Lovy and Friend 2015; Molnár et al. 2012; Whipps et al. 2012).

Due to their early-branching position within coccidians, *G. janae* and the related piscine coccidia may share numerous ancestral features. Indeed, two-valved sporocysts of *Goussia* spp. with a simple longitudinal suture seem to represent the ancestral state of the Sarcocystidae + Eimeriidae + Calyptosporidae clade, from which radiations into an array of excystation structures occurred (Jirků et al. 2002). These include the four-valved structures of sarcocystids, hemivalved arrangements with a thin membrane-covered oblong apical opening of the sporocyst wall of calyptosporids, and finally the univalved sporocysts containing Stieda and sub-Stieda bodies of eimerids (Whipps et al. 2012). A similar scenario can be envisioned for the evolution of host-parasite interactions, with the ancestral epicellular location retained in *G. janae* and related species parasitizing cold-blooded vertebrates, followed by a radiation of coccidians with various derived intracellular positions. However, the basal position of *G. janae* and its relatives may be affected by long-branch attraction and/or on phylogenies based on just 18S rRNA gene (Jirků et al. 2009). Hence, to resolve the above-mentioned uncertainties, it is important to extend the phylogenetic analyses by inclusion of protein-coding genes.

As every host and host tissue represent a nutritionally different environment, development of specific adaptations related to feeding behavior is a genuine part of the parasitic lifestyle. Indeed, when compared with the intracellular *Plasmodium* spp., the genetic analyses of epicellular

Cryptosporidium parvum and *C. hominis* have revealed the presence of several novel classes of cell-surface and secreted proteins with potential roles in host-parasite interactions and pathogenesis (Abrahamsen et al. 2004; Mazurie et al. 2013; Xu et al. 2004). However, a more extensive comparative analysis has been hindered by the lack of genomic and/or transcriptomic data from an epicellular coccidium.

In order to rectify this situation, we have performed morphological investigation of all life cycle stages of *G. janae*, and obtained the transcriptome of its sporulated phase which was selected because of the easy accessibility of a huge number of oocysts from host's casts minimally contaminated by host tissue. The collected data was used for extensive phylogenetic analyses and for predictions of the metabolic and invasive capabilities of this poorly molecularly characterized protist.

Results and Discussion

Morphology

In order to obtain sufficient quantity of starting material, we captured a dozen adult chubs in early spring when the temperature of the environment starts rising triggering the coccidium to develop and multiply in its host, thus reaching the highest infection densities during the year. While two fish were sacrificed to obtain material for light and electron microscopy, most were kept in an aquarium until they started to shed feces containing unsporulated oocysts. Heavily infected fish discharged white casts full of sporogonial stages of *G. janae* (and remnants of host epithelial cells), which started to sporulate after one week (Fig. 1A and B). Exogenous sporulation commences with the detachment of sporont cytoplasm from a thin oocyst wall (Fig. 1B; third from the top), cytoplasmic division into four globular or broadly oval sporoblasts (Fig. 1B; top), and final development into sporocysts, each containing two sporozoites (Fig. 1B; bottom). The morphology of the oocysts and sporocysts matched the original description of *G. janae* (Lukeš and Dyková 1990).

In the early phase of infection, the fish intestine was densely covered with small meronts that appeared to be attached to the microvillar surface of epithelial cells (Fig. 1C). In the more advanced phase of infection, large meronts, either undifferentiated or containing merozoites, as well as gamonts of both sexes were found in histological sections (Fig. 1D).

As revealed by transmission electron microscopy, all merogonial, gamogonial, and early sporogonial stages were intracellular, localized within the PVM squeezed between the enterocytic membrane and cytoplasm, in what is termed the epicellular localization. The earliest developmental stages observed were small dense granules and microneme-containing oval meronts, wedged among the microvilli of the enterocyte (Fig. 1G). Mature meronts had already developed typical apicomplexan features such as rhoptries, micronemes, and dense granules (Fig. 1F, J). Merozoites were usually formed by ectomerogony following multiple nuclear divisions and invaginations of meront's cytoplasm (Fig. 1E).

Early developmental stages assumed the monopodial position, with one extensive contact zone with the host cell (Fig. 1G, I, J). Later stages represented by quite large meronts, and macro- and microgametocytes were as a consequence of their growth extruded above the microvillar region. The subsequent growth of multiple thin projections that fused with the microvillar membrane (Fig. 1F, H) resulted in a "spider-like" attachment of the parasite (Fig. 1F, J). These are in fact fusions of heavily extended and distant regions of the same enterocyte membrane, apparently triggered by *G. janae* (Fig. 1K). The morphology of sexual stages was similar to that of other coccidians as previously described (Lukeš and Starý 1992). Large round macrogametocytes were observed, containing numerous dense granules, amylopectin, lipid inclusions, an apicoplast, an enlarged nucleus and endoplasmic reticulum (Fig. 1I), whereas numerous flagellated microgametes bud off from the microgametocytes' periphery (Fig. 1J, K).

Transcriptomic Analysis

In order to extend the morphology-based determination of the species, universal eukaryotic primers were used to amplify ~1,800 bp-long region of 18S rRNA gene, sequencing of which confirmed the appurtenance of the coccidian in question to *G. janae*.

Next, total RNA extracted from exogenously sporulated oocysts was DNase treated and its integrity was confirmed using Agilent Bioanalyzer 2100 (data not shown). In order to remove rRNA, the sample was subjected to poly-A selection and used for cDNA library preparation, which was paired-end Illumina sequenced on two lanes of the flow-cell to enhance sequence coverage. After initial quality checks on the reads, a total of 349,028,400 and 346,719,676 reads with a

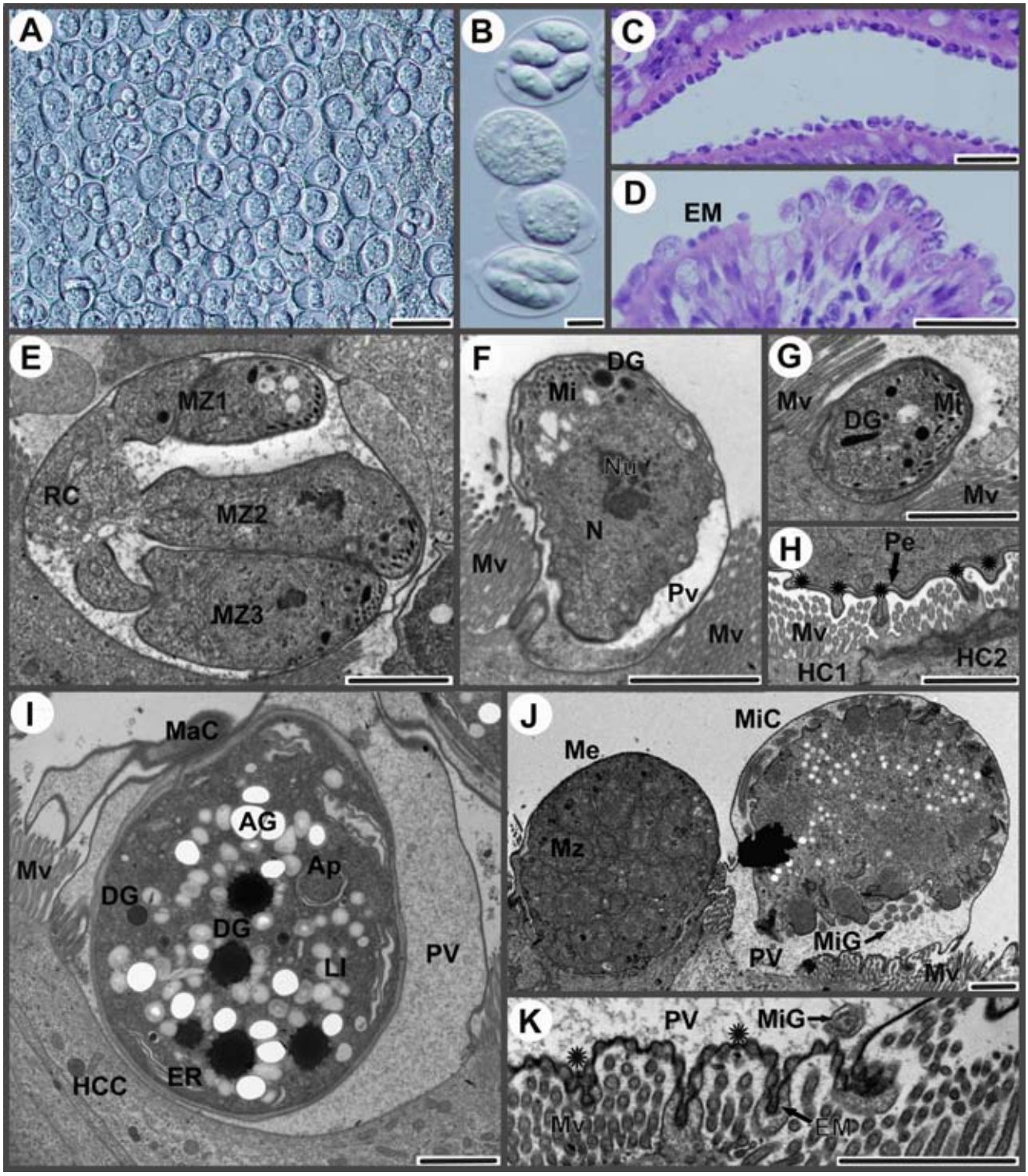

Figure 1. Light microscopy, histology and electron microscopy of developmental stages of *Goussia janae*. **A** and **B**. Fresh preparations of sporogonial stages of *G. janae* obtained from feces of *Squalius cephalus* and observed with differential interference contrast. **A**. Purified sporulating stages used for isolation of total RNA. **B**. Oocysts of *G. janae* at various stages of sporulation. Fully sporulated oocyst contains four sporocysts, each with two sporozoites. **C** and **D**. Haematoxylin and eosin-stained sections of the intestinal epithelium of adult *S. cephalus*. **C**. Early phase of the infection by *G. janae*, when the intestine is densely covered with the merogonial stages. **D**. Advanced infection with various developmental stages; EM – early meronts.

Table 1. Output of the Illumina HiSeq Sequencing run.

Sample	Index	Fragment size (bp)	Cycle Nb	Run Type
Goussia janae	ACAGTG	316	100	Multiplexed paired-end reads
	Yield (Mbases)	Raw Read Number	Mean Quality Score	Run Type
Lane 1	34903	349,028,400	35.71	Paired end
Lane 2	34672	346,719,676	35.71	Paired end

mean quality score of 35.71 were obtained, yielding 34,903 and 34,672 Mbp of sequences, respectively (Table 1) (data available in the ENA database; study accession number - PRJEB9672). The Trinity pipeline (Grabherr et al. 2011; Haas et al. 2013; Henschel et al. 2012) was used to assemble the filtered reads that generated 53,573 and 53,789 contigs. Their length distribution ranged from 200 bp to more than 10,000 bp with an average length of 774 bp and a mean GC content of 42% (Fig. 2). The most abundant raw sequences (over 80,000 contigs) were in the range of 200 to 1,200 bp.

Since the oocyst sample for RNA extraction had minute impurities from bacteria that grew during the week-long exogenous sporulation, as well as some RNA from the shed intestinal tissue of the fish host, the assembled transcripts were filtered for contamination by BLAST-alignment (Altschul et al. 1990) with high stringency onto the Swissprot protein database (Bairoch et al. 2004; Consortium 2015). The sequences corresponding to Cyprinidae (3,188) and various prokaryotes (331,679) were removed from subsequent analysis. Putative coding regions, at least 70 amino acids long, from the assembled Trinity transcripts were extracted using the TransDecoder utility, available in the Trinity software distribution, which yielded 44,742 peptide sequences.

A quick comparative analysis of the de novo assembled RNA-seq transcriptomes was performed using TRAPID (Van Bel et al. 2013), a software pipeline, which interrogates the transcripts against 'Reference proteome' databases, in this case the Alveolata clade database from OrthoMCL-DB (Chen et al. 2006). The output of the sequence similarity search contains assignments of each transcript to individual reference genes or gene families from closely related species. The best similarity search hits from the TRAPID analysis showed the highest sequence similarity to *T. gondii* and *N. caninum*, with 44.3% and 40.3% of the sequences assigned to them, respectively (Fig. 2).

Annotation of Consensus Sequences

A superficial analysis of the amino acid sequences was done using KAAS (KEGG Automatic Annotation Server)(Moriya et al. 2007), which provides functional annotation of genes by BLAST

E-K. Transmission electron microscopy of developmental stages of *G. janae* in the intestine of adult *S. cephalus*. **E**. Ectomerogonial division with merozoites (MZ1, MZ2, MZ3) protruding from residual cytoplasm (RC). **F**. Mature meront with micronemes (Mi), dense granules (DG), nucleus (N) and central nucleolus (Nu) surrounded by the parasitophorous vacuole (PV) and host cell microvilli (Mv). Via multiple projections of the enterocyte membrane, the early monopodial epicellular location is changing into the spider-like location. **G**. Early meront containing dense granules (DG) and micronemes (Mi) is wedged among the microvilli (Mv) of the enterocyte. **H**. Detail of the interface of spider-like merogonial stage extending projections (asterisks) into the microvilli (Mv) of two host's cells (HC1, HC2). **I**. Macrogametocyte (MaC) in mostly monopodial location, containing amylopectin granules (AG), lipid inclusions (LI), dense granules (DG), endoplasmic reticulum (ER) and apicoplast (Ap); HCC – host cell cytoplasm, Mv – microvilli, Pv – parasitophorous vacuole. **J**. Late merogonial stage (Me) with merozoite nuclei (Mz) and a microgametocyte (MiC) containing numerous microgametes (MiG) at its periphery; Pv – parasitophorous vacuole, Mv – microvilli. **K**. Detail of the host-parasite interface with multiple fusions (asterisks) of individual microvilli (Mv) and the enterocyte membrane (EM) covering the parasite; Pv – parasitophorous vacuole, Mv – microvilli. Scale bars: A = 50 μm; B = 13 μm; C, D = 20 μm; E–K = 2 μm.

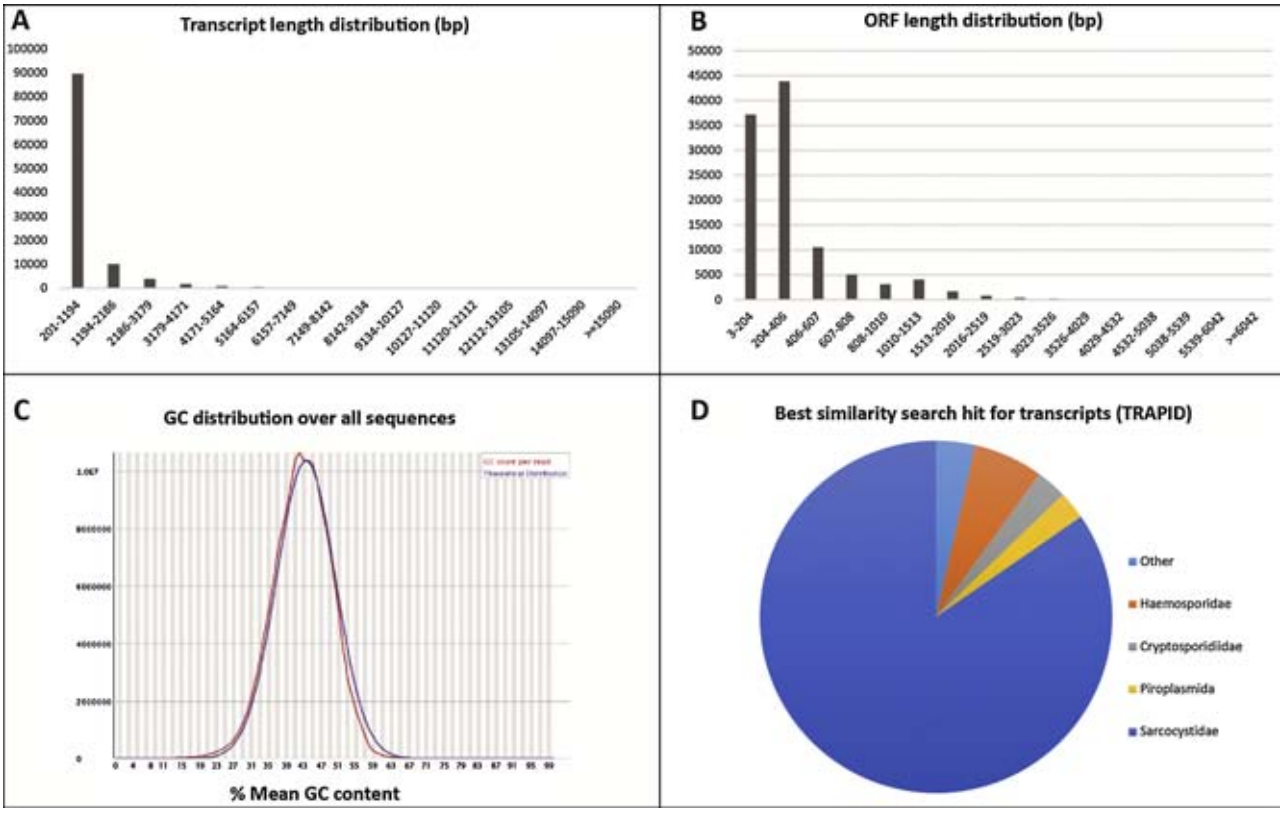

Figure 2. (**A** and **B**) The length distribution of the contigs and predicted ORFs obtained from the de novo assembly of high-quality clean reads by Trinity. (**C**) The GC distribution of the sequences across the Trinity assembled sequences (**D**) Pie-chart of the species distribution for the Trinity contig sequences mapped against the Alveolata clade database.

(Altschul et al. 1990) comparisons against the manually curated KEGG genes database. The result contains KO (KEGG Orthology) assignments and automatically generated KEGG pathways. The KO assignment was performed based on the bi-directional best hit (BBH) of BLAST against alveolates as the representative data set.

The output categorized the sequences into 251 groups, belonging to different cellular processes and functions. The results comprised a number of categories including "Biosynthesis of secondary metabolites", "Oxidative phosphorylation", "Glycolysis/Gluconeogenesis", "Biosynthesis of amino acids", "Amino sugar and nucleotide sugar metabolism", "Starch and sucrose metabolism", "Pentose phosphate pathway", "Fatty acid metabolism", "Terpenoid backbone biosynthesis", "Alanine, aspartate and glutamate metabolism", among others (Supplementary Material Table S1).

For a detailed annotation, the Trinotate suite (http://trinotate.github.io/) was employed, wherein the consensus sequences were first searched using BlastX and BlastP against the Swissprot (Bairoch et al. 2004) and EMBL databases (Kanz et al. 2005). The suite also searched for protein domains by HMMER (against PFAM database) (Finn et al. 2011; Punta et al. 2012), signal peptide (signalIP) (Petersen et al. 2011), and transmembrane domains (tmHMM) (Krogh et al. 2001). The BlastX and BlastP searches against Swissprot resulted in 4,913 and 3,758 annotations, while the search against the EMBL database yielded 5,571 and 8,631 annotations, respectively (Table 2 and Supplementary Material Table S2). Again, among the annotated sequences, the two coccidians with by far the highest number of best hits were *T. gondii* and *N. caninum* (Table 2).

Phylogenetic Analyses

Phylogenetic analyses based on protein sequences were done to assess the relationships of *G. janae* with other apicomplexans, for which sequences were available. The analyses of deduced amino acid sequences of *G. janae* together with other

Table 2. Tophits of the blast results of Trinity sequences of *Goussia janae* against SwissProt and EMBL databases.

Trinotate		Total	Alveolates	*Toxoplasma gondii*	*Neospora caninum*
SwissProt database	blastx	4913	634	174	3
	blastp	3758	400	129	2
EMBL database	blaspx	5571	3941	1622	1278
	blastp	8631	6883	2887	2330

Apicomplexa were based on glyceraldehyde-phosphate dehydrogenase (GAPDH), pyruvate dehydrogenase E1 beta (PDHE1 β), malate dehydrogenase, and 6-phosphofructokinase, all single-copy protein-coding genes known to be suitable for apicomplexan phylogenetic inference (Kuo et al. 2008). On the basis of its intestinal and epicellular localization, it would be logical to assume that *G. janae* is related to coccidians with similar tissue localization, namely to members of the genus *Cryptosporidium*. However, none of our analyses affiliated *G. janae* exclusively with cryptosporidians; only one tree has shown *G. janae* positioned at the base of eimeriorinids + cryptosporidians (MDH in Supplementary Material Fig. S3). Most of our phylogenetic analyses showed that *G. janae* clustered basal to the whole eimeriorinid clade or to sarcocystids (Fig. 3, and Supplementary Material Figs S3 and S4) which are in the sequence databases represented primarily by *T. gondii* and *N. caninum*, economically relevant parasites of warm-blooded vertebrates. Such relationships were previously also revealed by 18S rDNA data (Bartošová-Sojková et al. 2015; Jirků et al. 2002, 2009; Molnár et al. 2012; Whipps et al. 2012). Exceptionally, *G. janae* clustered at the base of the eimeriid subclade (GAPDH in Supplementary Material Fig. S4) or within the sarcocystid subclade (PDHE1B in Fig. 3). The results of our phylogenetic analyses indicate that the epicellular parasitism arose independently in *G. janae* and cryptosporidians that is in agreement with the interpretations based on a large set of 18S rDNA data (Bartošová-Sojková et al. 2015).

Analyzing *G. janae* Transcripts by BLAST-alignment onto *T. gondii*

Since the species distribution from the TRAPID analysis showed that the *G. janae* transcripts share the highest similarity with *T. gondii* and *N. caninum*, a blast search (p-value 10e-7) was performed against their genomes to search for and deduce the presence of corresponding genes in *G. janae*. The transcripts mapped to 3,546 and 3,411 genes of *T. gondii* and *N. caninum*, respectively (Supplementary Material Table S3), as annotated in ToxoDB. Based on these results, the contigs and the corresponding genes were checked for their presence in metabolic pathways, as annotated in ToxoDB (Gajria et al. 2008) and www.llamp.net (Shanmugasundram et al. 2013).

The contigs were analyzed independently and manually, comparing them to the corresponding blast results (against *T. gondii*), and were consequently assigned a score based on the amount of coverage of proteins detected and tabulated as a file (Tables 3, 4, Supplementary Material Tables S4 and S5). Some genes have several contigs mapped to them, which is likely due to low coverage. In case of insufficient overlap and read coverage, the assembler fails to join contigs, resulting in several different contigs matching a single transcript. Another possible explanation for this situation is alternative splicing, which, however, was not analyzed herein. Finally, some genes might have been missed either due to low/no expression at the investigated oocyst/sporozoite stage – these genes were indicated as n.d (not detected) in the tabulated list. It should also be noted that the presence of transcripts might not directly translate to protein and a functional enzyme at this stage.

Central Carbon Metabolism

It is vital to understand parasite metabolic capabilities and host-parasite interaction in order to delineate the establishment of intracellular parasitism of apicomplexans over the course of evolution (Danne et al. 2013; Fleige et al. 2010; Seeber et al. 2008). The apicomplexans encounter different niches during their different life cycle stages and they need to evolve flexible modes of nutrient acquisition. Apart from its own metabolic capabilities, the epicellular location of *G. janae* provides two likely nutrient sources - the host cytosol and the lumen of the gut. We sought to understand the metabolic capabilities of *G. janae* at the

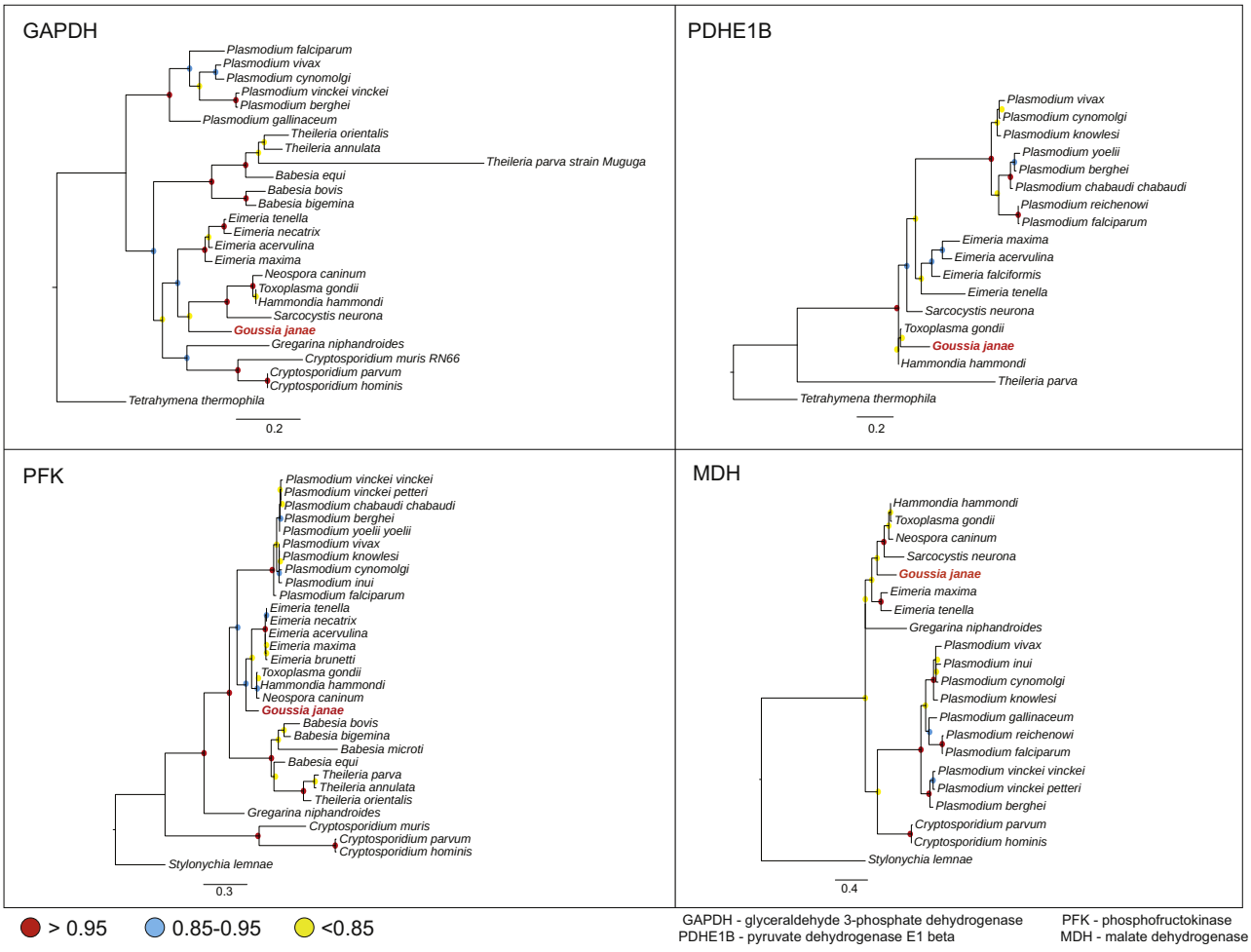

Figure 3. Maximum likelihood phylogenetic trees based on *G. janae* protein sequences predicted from de-novo assembled Trinity transcripts and corresponding protein sequences from other apicomplexan organisms. The tree was constructed using PhyML using WAG model of amino acids substitution with NNI topology search, based on an amino acid alignment by MUSCLE. Ciliate protein sequences (from *Tetrahymena thermophila* or *Stylonychia lemnae*) were used as outgroups. The SH-like aLRT branch support values were indicated with colored discs: red >0.95, blue 0.85-0.95 & yellow <0.85. The scale bar at the base of each phylogenetic tree represents the branch length values, the number of substitutions per site.
PFK – phosphofructokinase; GAPDH – glyceraldehyde 3-phosphate dehydrogenase; MDH – malate dehydrogenase; PDHE1B – pyruvate dehydrogenase E1 beta subunit.

investigated oocyst/sporozoite stage, by analyzing the enzymes that could be detected in the assembled transcriptome. The detected repertoire of the metabolic enzymes suggests an active metabolic state in the sporulated oocysts of *G. janae*. However, given the transitionary nature of the stage being investigated, it is difficult to conclusively ascertain the metabolic capabilities of the parasite.

Glucose and other sugars should be either taken up from the host, produced by gluconeogenesis or by degrading amylopectin (Coppin et al. 2003). According to the analysis of the transcriptomic data, *G. janae* possesses a glycogen debranching enzyme and glycogen phosphorylase, suggesting that it can store and mobilize glucose from amylopectin. Concordantly, morphological analysis showed the presence of amylopectin granules in the macrogametocyte stage (Fig. 1I). Analysis of the starch metabolism pathway suggests that the studied coccidium might be capable of utilizing an UDP-glucose-based pathway to synthesize starch, as suggested for other Apicomplexa (Supplementary Material Table S4F). Genes homologous to UDP-glucose 4-epimerase, glycogen (amylopectin/

Table 3. List of detected homologues of the invasion and glideosome machinery relative to *T. gondii*.

Glideosome associated proteins	Homologue in TGME49	Seq. coverage	Orthologue in CryptoDB
Myosin A	TGME49_235470	~100%	Yes
Myosin B/C,	TGME49_255190	~100%	Yes
Myosin light chain, MLC1	TGME49_257680	~100%	Yes
Essential Light chain, ELC1	TGME49_269438	n.d	No
GAP40	TGME49_249850	~100%	Yes
GAP45	TGME49_223940	~100%	No
GAP50	TGME49_219320	~100%	Yes
GAP70	TGME49_233030	n.d	No
GAP80	TGME49_246940	n.d	No
GAPM1a	TGME49_202500	n.d	Yes
GAPM1b	TGME49_202510	n.d	Yes
GAPM2a	TGME49_219270	~100%	No
GAPM2B	TGME49_206690	n.d	Yes
GAPM3	TGME49_271970	n.d	Yes
actin	TGME49_209030	~100%	Yes
Moving Junction	**Homologue in TGME49**	**Seq. coverage**	**Orthologue in CryptoDB**
apical membrane antigen, AMA1	TGME49_255260	>90%	No
AMA2	TGME49_300130	~30%	No
AMA3 or spAMA1	TGME49_315730	>50%	No
AMA4	TGME49_294330	>80%	No
RON2	TGME49_300100	~100%	No
RON4	TGME49_229010	~33%	No
RON5	TGME49_311470	~50%	No
RON8	TGME49_306060	~60%	No
spRON2 or $RON2_{L2}$	TGME49_265120	~100%	No
$RON2_{L1}$	TGME49_294400	n.d	No
Conoid	**Homologue in TGME49**	**Seq. coverage**	**Orthologue in CryptoDB**
apical complex lysine methyltransferase (AKMT)	TGME49_216080	n.d	Yes
calmodulin CAM1 (CAM1)	TGME49_246930	~100%	No
calmodulin CAM2 (CAM2)	TGME49_262010	n.d	Yes
Myosin H	TGME49_243250	~100%	Yes
RNG1	TGME49_243545	n.d	No
RNG2	TGME49_244470	n.d	No
centrin 2	TGME49_250340	~100%	Yes
dynein light chain DLC	TGME49_223000	~100%	Yes
MICs/Thrombospondin related anonymous protein (TRAP) family	**Homologue in TGME49**	**Seq. coverage**	**Orthologue in CryptoDB**
microneme protein 2 (MIC2)	TGME49_201780	n.d	No. Has other TRAP domain-containing proteins
MIC2 associated protein, M2AP	TGME49_214940	n.d	No
microneme protein MIC1	TGME49_291890	n.d	No

Table 3 (*Continued*)

MICs/Thrombospondin related anonymous protein (TRAP) family	Homologue in TGME49	Seq. coverage	Orthologue in CryptoDB
microneme protein MIC4	TGME49_208030	n.d	No
microneme protein MIC6	TGME49_218520	n.d	No
microneme protein MIC3	TGME49_319560	n.d	No
microneme protein MIC8	TGME49_245490	n.d	No
microneme protein MIC14 (TRAP domain containing protein)	TGME49_218310	>50%	No
microneme protein MIC15 (TRAP domain containing protein)	TGME49_247195	~66%	No
microneme protein MIC16 (TRAP domain containing protein)	TGME49_289630	~40%	No
microneme protein MIC12 (MIC12)	TGME49_267680	~100%	Yes
microneme protein MIC7 (MIC7)	TGME49_261780	~100%	No
sporozoite protein with an altered thrombospondin repeat SPATR	TGME49_293900	~50%	No
alpha-tubulin suppressor protein	TGME49_267450	~100%	No
thrombospondin type 1 domain-containing protein	TGME49_277910	~50%	No
PAN/Apple domain-containing protein	TGME49_200270	~100%	No
DJ-1 family protein	TGME49_214290	>75%	Yes
Inner membrane complex & associated components	**Homologue in TGME49**	**Seq. coverage**	**Orthologue in CryptoDB**
IMC sub-compartment protein ISP1	TGME49_260820	~100%	Yes
IMC sub-compartment protein ISP2 (ISP2)	TGME49_237820	~100%	Yes
IMC sub-compartment protein ISP3 (ISP3)	TGME49_316540	>80%	No
IMC sub-compartment protein ISP4 (ISP4)	TGME49_205480	n.d	No
IMC1 (ALV1)	TGME49_231640	100%	No
IMC3 (ALV3)	TGME49_216000	~60%	Yes
IMC4 (ALV4)	TGME49_231630	~100%	Yes
IMC5 (ALV11)	TGME49_224530	~40%	No
IMC6 (ALV6)	TGME49_220270	~80%	Yes
IMC7 (IMC7)	TGME49_222220	n.d	No
IMC8 (ALV10)	TGME49_224520	~80%	No
IMC9 (ALV6)	TGME49_226220	~50%	Yes
IMC10 (ALV12)	TGME49_230210	~100%	Yes
IMC11 (ALV7)	TGME49_239770	n.d	No
IMC12 (ALV12)	TGME49_248700	~50%	No
IMC13 (ALV8)	TGME49_253470	>85%	Yes
IMC14 (ALV9)	TGME49_260540	~33%	No
IMC15 (ALV5)	TGME49_275670	n.d	No
inner membrane complex protein IMC2A (IMC2A)	TGME49_228170	>90%	No
inner membrane complex protein IMC3 (IMC3)	TGME49_244030	n.d	No

Table 4. List of detected homologues of dense granule proteins corresponding to *T. gondii*.

Dense granule proteins	Homologue in *T. gondii*	Seq. coverage	Orthologue in CryptoDB
dense granular protein GRA10 (GRA10)	TGME49_268900	~55%	No
dense granule protein GRA12 (GRA12)	TGME49_275850	~80%	No
dense granule protein GRA12 (GRA12) or GRA12 bis	TGME49_288650	~100%	No
dense-granule antigen DG32 (GRA17)	TGME49_222170	~60%	No
dense granule protein DG32 or GRA23	TGME49_297880	n.d	No
mitochondrial association factor - 1	TGGT1_220950	n.d	No
NTPase	TGME49_225290	~40%	No
nucleoside-triphosphatase	TGME49_277720	>70%	No
NTPase I	TGME49_277240	n.d	No
NTPase II	TGME49_277270	n.d	No
protease inhibitor PI1 (PI1)	TGME49_217430	n.d	No
protease inhibitor PI2 (PI2)	TGME49_208450	n.d	No
serine proteinase inhibitor PI-2, putative	TGME49_208430	n.d	No
cathepsin CPC1 (CPC1)	TGME49_289620	~90%	Yes
cathepsin CPC2 (CPC2)	TGME49_276130	~100%	Yes
cyclophilin (Cyclophilin-18)	TGME49_221210	n.d	No
14-3-3 protein	TGME49_263090	~75%	Yes
Other dense granule proteins GRA1-9,11,14-16,19-25	-	n.d	No

floridean starch) synthase, bifunctional trehalose phosphate synthase/trehalose phosphatase, and 1,4-alpha-glucan branching enzyme among others were detected.

G. janae possesses almost all components of the classical glycolytic cycle (Supplementary Material Table S4A), leading to the formation of phosphoenolpyruvate (PEP) from glucose-6-phosphate. The PEP thus formed can either be converted to pyruvate, by pyruvate kinase, or acted upon by phosphoenolpyruvate carboxylase to form oxaloacetate. However, we could detect neither of these enzymes in the *G. janae* transcriptome. Besides, the first enzymes of the pathway, hexokinase and aldolase, could also not be detected at the investigated stage. The absence of hexokinase may suggest that the oocysts/sporozoites of *G. janae* make use of stored amylopectin, degrading it to glucose-1-phosphate via glycogen phosphorylase and initiating glycolysis with phosphoglucomutase activity, a situation reminiscent of *C. parvum* (Entrala and Mascaró 1997). The amylopectin needed for fueling the developmental stage transition and the initial infective sporozoites might be synthesized and stored at the macrogametocyte stage, where enzymes that were not detected at the investigated stage might be expressed. The absence of pyruvate kinase could be explained by its regulation by glucose-6-phosphate, a product of hexokinase, which could not be detected at this stage of the parasite (Saito et al. 2008).

The pyrophosphate-dependent 6-phosphofructokinase (pyrophosphate: d-fructose 6-phosphate 1-phosphotransferase; PFK) was detected, indicating that the parasite relies on it, instead of ATP-PFK in this stage. *Cryptosporidium* species possess one lactate dehydrogenase, which converts pyruvate to lactate, whereas *T. gondii* possesses two isoforms of this enzyme (Ferguson et al. 2002; Yang and Parmley 1997), one of which is detected in this stage of *G. janae.*

Within the transcriptome, we have detected a complete set of enzymes required for the TCA cycle (Supplementary Material Table S4B), indicating the capability of the studied parasite to utilize acetyl-CoA from either oxidation of pyruvate or fatty acid/branched-chain amino acid degradation

as the enzymes corresponding to this pathway were detected as in *T. gondii* (Limenitakis et al. 2013). The genes detected indicate that *G. janae* could potentially oxidize glucose, various amino acids, and fatty acids via the TCA cycle. Interestingly, genes for all enzymes of this pathway are present in the *C. muris* genome, but are absent in *C. parvum* and *C. hominis* (Xu et al. 2004) *C. muris* possesses one citrate synthase as opposed to three isoenzymes found in *T. gondii*, two of which (TGME49_203110 and TGME49_268890) have orthologues in *G. janae*.

The pentose-phosphate pathway (PPP), or hexose-monophosphate pathway, is a source of NADPH for both biosynthetic purposes and oxidative stress protection, while at the same time produces ribose moieties for the synthesis of nucleic acids. Majority of the enzymes of the PPP including sedoheptulose-1,7-bisphosphatase were detected, indicating the capacity to utilize ribose as an alternate or supplementary carbon source (Supplementary Material Table S4C). *G. janae* also expresses most of the genes involved in the branched mitochondrial respiratory chain at this stage suggesting that it is capable of oxidative phosphorylation.

The dataset obtained from BLAST alignment of the *G. janae* transcripts onto *T. gondii* protein sequences (ToxoDB) (Gajria et al. 2008) consists of 1,416 genes. This gives a general view of the protein repertoire available to *G. janae*. Within this, genes involved in the metabolism of pyruvate, glutamate and fatty acids were detected, suggesting that they can be utilized by the parasite for energy production. Transcripts coding for the enzymes involved in the fatty acid synthesis pathways I and II (FAS I and II) were detected along with an almost complete set of genes involved in the elongation of fatty acids via elongase pathway of endoplasmic reticulum. Though we could not determine the presence of the bipartite signal motif due to partial sequence information, the presence of a typical apicoplast (Fig. 1I) and detection of a few apicoplast enzymes of the FASII system suggest the existence of other enzymes allowing generation of fatty acids and isoprenoids.

Gliding Motility, Host Cell Attachment and Invasion

Host cell entry by most apicomplexans is an active, multistep process requiring recognition and attachment of the host cell, followed by the formation of a so-called moving junction (MJ) with the host cell and subsequent penetration (Carruthers and Boothroyd 2007). This active invasion of host cells is mostly dependent on the machinery of gliding motility, termed glideosome, which is conserved across the Apicomplexa.

In *T. gondii*, motility is driven by rearward translocation of adhesion complexes (formed by the microneme proteins (MICs) inserted in the parasite plasma membrane, and their receptors on the host cell) by the action of the actomyosin system located underneath the plasma membrane. During invasion, the parasite sequentially discharges a series of apical secretory organelles (micronemes, rhoptries) to enact efficient gliding motility. Myosin A (TgMyoA) translocates adhesins at the MJ from the apical to the posterior pole of the parasite. The MJ is composed of apical membrane antigen 1 (AMA1) associated with a complex of rhoptry neck proteins (RONs) that are anchored into the host cortical cytoskeleton, and serves as a support for the parasite's propulsion into the host cell (Tyler et al. 2011). Invasion occurs rapidly within 30-45 seconds, upon which time the coccidium resides within a parasitophorous vacuole (PV) formed during penetration from both host and parasite material.

The assembled *Goussia* transcriptome was searched for genes associated with the glideosome, attachment (adhesins) and invasion machinery (components of the MJ) relative to *T. gondii*, and tabulated (Table 3). *G. janae* attachment and invasion components are broadly conserved with those in *T. gondii*, involving a set of proteins that coordinate with the parasite and host. Besides the stage specificity and RNA-seq coverage, the absence or non-detection of certain invasion proteins in comparison to *T. gondii* could be explained by their sequence divergence and host-specific evolution. The search for invasion proteins identified a few orthologues of *T. gondii* microneme proteins (MIC7, MIC12). We also identified potential orthologues of MIC13, MIC14, MIC15 and MIC16 (contigs' ORF spanning ~50% of the *T. gondii* protein sequences) as well as orthologues of PAN/Apple domain-containing protein (TGME49_200270, TGME49_286150) (Lamarque et al. 2014) and sporozoite proteins with an altered thrombospondin repeat SPATR (TGME49_293900). Thrombospondin-related anonymous protein (TRAP) domain containing proteins in *T. gondii* and *Plasmodium* spp. (TgMIC2 and TRAP) transmit the force generated by the actomyosin motor to the host cell (Carruthers and Tomley 2008). The C-terminal tail of these proteins is assumed to be linked to an actin filament via bridging molecules. Myosins anchored to the

inner membrane complex (IMC) pull on the actin filament creating the necessary locomotive force to drive forth the parasite (Boucher and Bosch 2015). However, no homologue of TgMIC2 was detected, implicating an important role for other TRAP domain-containing proteins or a potential new factor.

G. janae possesses homologues of actin and several glideosome components of *T. gondii*, including myosin A, myosin light chain 1 and actin-depolymerizing factor (ADF). The actomyosin motor is attached to the IMC, a critical organelle required for host invasion, and is comprised of flattened alveolar sacs. The transcriptome analysis identified putative orthologues of IMC1, IMC4, IMC6, IMC8, IMC10, IMC13 as well as sequences spanning 30-60% of the sequences of IMC3, IMC5, IMC9, IMC12, IMC14, IMC20 and IMC24 (Anderson-White et al. 2011; Chen et al. 2015). *G. janae* possesses the IMC sub-compartment proteins ISP1-3, but not ISP4 (Beck et al. 2010). The gliding-associated proteins, GAP40, GAP45, and GAP50, anchoring the actomyosin motor to the IMC (Frénal et al. 2010; Gaskins et al. 2004), are also well conserved in *G. janae*, however GAP70, GAP80 and GAPM proteins are lacking. Concordantly with the absence of GAP80, the myosin located to the posterior pole, MyoC was found (Frénal et al. 2014).

The apical membrane antigen 1 (AMA1) links the inner membrane complex of the parasite to the host cell via interactions with rhoptry neck proteins (RONs), forming the AMA1-RON2-4-5-8 complex, that together make up the MJ (Alexander et al. 2005; Besteiro et al. 2009; Tyler et al. 2011). AMA1 is broadly conserved across the Apicomplexa, however interestingly, *Cryptosporidium* lacks an AMA1 orthologue. In addition to AMA1, *G. janae* also appears to possess the two additional AMA1 paralogs (TGME49_300130 and TGME49_315730) corresponding to AMA2 and sporoAMA1/AMA3 recently described (Lamarque et al. 2014; Poukchanski et al. 2013). Additionally, RONs were found to be well conserved in *G. janae* and *T. gondii*, with RON2 and RON3 orthologues displaying complete sequence coverage and significant sequence similarity. Putative *G. janae* orthologues of RON8, RON9 and RON10 were identified as well, albeit with lesser sequence coverage. Also, predicted ORFs that could be aligned over 25-50% sequences of RON1, RON4, RON5 and RON6 were detected. The presence of MJ components in *G. janae* is rather unexpected and contrasts with the epicellular cryptosporidians that are lacking them.

T. gondii secretes rhoptry and dense granule proteins (GRAs) that interact with the PVM and host cell targets, manipulating pathways that protect the intracellular parasite against clearance. In contrast to the invasion machinery, proteins implicated in the subversion of host cellular functions are predominantly species-specific and those characterized recently in *T. gondii* are not found in *G. janae*. Only ROP9 homologue, an early invasive stage protein was detected (Chen et al. 2014). Among the GRAs, homologues of GRA10, GRA12, GRA12 bis, GRA17 and a few other putative GRAs-like cathepsins (CPC1 and CPC2) were detected (Table 4). Importantly, *T. gondii* GRA17 and GRA23 have recently been demonstrated to localize to the PVM and contribute to the molecular sieve mediating passive transport of small molecules across the membrane (Gold et al. 2015). Most apicomplexans that form a PMV concomitantly possess either one or both these GRAs providing for PVM permeability and nutrient flux. However, *Cryptosporidium* spp. lack either of them, suggesting that the molecular sieve is replaced in these parasites by the feeder organelle membrane (FOM) as the major site of nutrient transport. GRA12 is associated with the intravacuolar membranous nanotubular network, which contributes to the membranous interface between the parasite and host cell (Besteiro et al. 2009). The non-detection of other GRA proteins could be explained by their absence in the investigated stage or by the fact that the majority of the secreted proteins encoded by *T. gondii* target host signaling pathways, and a different set of proteins are involved in modulating the fish host. Furthermore, the epicellular localisation might make the parasite less exposed to the host cytosol and thus requires a different or smaller repertoire of proteins.

The necessary proteins to drive its entry into the host cell by actin-dependent gliding motility, as described for *T. gondii* and *Plasmodium* spp., have also been found in *G. janae*. However, this coccidium does not penetrate deep into the cytosol of the host cell, instead residing epicellularly beneath the cell membrane in the apices of intestinal epithelial cells in a manner similar to members of the genus *Cryptosporidium*. The microvilli of the intestine are stretched around the parasite and the epithelial membrane is fused, implicating a role for modulation of the underlying host cell cytoskeleton following infection. This is in contrast to *T. gondii*, which upon infection in the gut traverses the epithelial monolayer, crosses the basement membrane, ultimately reaching the circulatory system. Despite the presence of important proteins involved

in the glideosome and invasion machinery, relative to *T. gondii*, the restricted epicellular localisation of *G. janae* and inability to completely invade the host cell, suggests alternate factors are involved. In this context the absence of the MyoC-glideosome (Frénal et al. 2014) and the members of GAPMs family (Bullen et al. 2009) as well as some components of the conoid such as RNG 1, RNG2 and AKMT (Katris et al. 2014; Sivagurunathan et al. 2013) might reflect the different needs in host cell entry and egress.

Conclusions

Results of mRNA sequencing and phylogenetic and BLAST analyses are in agreement with the previous 18S rDNA-based observations of the basal clustering of *G. janae* in the eimeriorinids or sarcocystids. All things considered, the abundant well-curated data sets for the Sarcocystidae could potentially skew this inference in their favor.

G. janae genes involved in the glideosome function and MJ formation relative to *T. gondii* were identified. *G. janae* seems able to rely on the conserved actomyosin motor for motility and penetration, a well conserved feature of apicomplexan parasites studied thus far. However, it resides in an epicellular location at the apical surface of intestinal epithelial cells, enclosed by the host cell plasma membrane, reminiscent of *Cryptosporidium*. The results from this search highlight the potential importance of conserved invasion machinery proteins. *G. janae* is well positioned to help answering how epicellular parasites might have evolved to become obligate intracellular parasites. It is intriguing to investigate further the evolution of this coccidium in metabolically adapting to these two different niches. Its epicellular location constrains its contact to the host cytosol although the parasite is likely accessing host cell nutrient via the molecular sieve at the PVM. Moreover, *G. janae* is in close proximity to the gut lumen, another nutrient source, and harbors comparable metabolic capabilities compared to *T. gondii* indicative of a potentially high level of versatility.

A comprehensive transcriptomic analysis performed on a series of time points covering the entire oocyst development in *T. gondii* (Fritz et al. 2012) has provided important insights into sporogenesis. The data reported here on *G. janae* should contribute to the identification of the conserved genes governing oocyst wall formation and resistance to environment as well as sporozoite infectivity.

Methods

Oocysts source and purification: 35 chub (*Squalius cephalus*) fish were sampled with electrofishing method in Plav on Malše River, Czech Republic in March 2014 and kept in the aquaria at 8 °C. The fish were starved to increase the release of as much clean casts as possible. Two days after sampling, fish started to release white casts. The light microscopic observation of the casts revealed they contain *Goussia janae*-like oocysts, the species previously described from common dace and chub (Lukeš and Dyková 1990). As most of the casts contained only a small amount of oocysts, fish were moved to room temperature (20 °C) to increase the oocyst production. The casts were collected every day and placed in separate Petri dishes maintained at 8 °C room to sporulate. Two casts containing a large amount of oocysts were randomly selected for obtaining 18S rDNA sequences for the isolate's characterization. The casts were left to sporulate for two weeks. 65 casts released from *S. cephalus* were examined under the Olympus BX51 light microscope to observe the sporulation process, as well as presence of contaminants (bacteria, ciliates, diatoms, trematode eggs, myxozoans-*Myxobolus*). Potassium dichromate was added to *S. cephalus* casts to eliminate the bacteria and ciliates as well as to prevent their further growth. The oocysts were purified from the host tissue by centrifugation in a cesium chloride gradient, and ampicillin and penicillin were added to prevent the growth and eliminate the bacteria, albeit incompletely. Photographic documentation was done during oocyst sporulation using differential interference contrast microscopy on the Olympus BX51 light microscope equipped with Olympus DP70 digital camera. The sample did not contain any myxozoan spores, trematode eggs or diatoms except for some host tissue, which was not possible to remove. The sample also contained some bacteria, which were not eliminated after antibiotic treatment.

Histology and electron microscopy: Parts of the intestinal tissue were fixed, processed and stained for histology as described previously (Lukeš and Dyková 1990). For transmission electron microscopy, small parts of the infected intestine were fixed in 4% OsO_4 in 0.2 M cacodylate buffer (pH 7.2) for 1 hr at 4 °C, then dehydrated in graded ethanol series and embedded in Epon-Araldite resin, and ultrathin sections were viewed in a JEOL JEM-1010 transmission electron microscope. Fecal casts from fish kept in aquaria were carefully collected by a Pasteur pipette and transferred into a Petri dish containing tap water, where they were kept to sporulate at 15 °C for 1 week. Next, the oocysts were collected by differential centrifugation and a sample containing approximately 6,050,000 oocysts was used for RNA isolation (see below).

RNA isolation: The homogenized sample was lysed in TRIzol (Life Technologies), incubated for 5 min at RT, and the isolation was performed by chloroform-isopropanol extraction according to manufacturer's instructions (Life Technologies). The RNA pellet was air dried for 5 min and then resuspended in RNase-free water, incubated at 55 °C for 10 min and stored at -80 °C until use. Total RNA quality and concentration was assessed and subsequently sent for RNA sequencing. The sample yielded 1.741 μg of RNA when eluted in 50 μl of H_2O, at a concentration of 34.82 ng/μl (Supplementary Material Fig. S2).

DNA extraction from casts, PCR amplification and sequencing for strain identification: Total DNA from two randomly selected casts was extracted by a standard phenol-chloroform method) after an overnight digestion with proteinase K (100 μg ml^{-1}) at 55 °C (Maslov et al. 1996). The extracted DNA was re-suspended in 100 μl of distilled H_2O and kept at

4 °C. 18S rRNA gene was amplified from both samples using universal eukaryotic primers ERIB1 and ERIB10 (Barta et al. 1997). The PCR program consisted of initial denaturation, followed by 30 cycles of 95 °C for 1 min, 48 °C for 1 min, 72 °C for 2 min. The PCR products were purified and sequenced.

Library preparation and Illumina paired-end RNA-seq: Illumina HiSeq 2500 paired-end (2x100 bp) library preparation and sequencing was carried out by the Genomics platform, iGE3 (the Institute of Genetics and Genomics) at the University of Geneva. The RNA sample was multiplexed across two sequencing lanes of the flow cell.

De novo transcriptome assembly by Trinity: The adapter sequences from the raw reads were trimmed using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) (phred<20). The resulting reads after quality control were fed into the Trinity pipeline for *de novo* assembly (Grabherr et al. 2011; Haas et al. 2013; Henschel et al. 2012), performed on the Baobab cluster of the University of Geneva. The read assembly was performed with a minimum k-mer coverage of 2 and the default k-mer size of 25. The Trinity modules assemble the RNA-seq reads into full-length transcripts, and has been found to be efficient in *de novo* transcriptome reconstruction, better than other available de novo transcriptome assemblers.

Annotation: The assembled sequences were used for blast searches against a number of protein databases such as Swiss-Prot, UniRef90, OrthoMCL-DB (TRAPID pipeline) and KEGG.

KAAS (KEGG Automatic Annotation Server), was used for a quick functional annotation of the Trinity transcripts, to get a broad overview of the functional classification of the transcripts (Kanehisa et al. 2012; Moriya et al. 2007). KAAS blasts the transcripts to the manually curated KEGG GENES database and assigns the corresponding KEGG orthoglogy identifiers (K numbers) to genes in the KEGG pathway database.

The Trinotate suite was used for a comprehensive annotation of the Trinity assembled sequences (http://trinotate.github.io/). Trinotate is designed for automatic functional annotation of transcriptomes, where Trinity transcripts are searched for sequence homologies using BLAST (Altschul et al., 1990) against available sequence data (SwissProt/Uniref90). The suite includes protein domain identification by HMMER (PFAM) (Finn et al. 2011; Punta et al. 2012), signal peptide and transmembrane domain prediction (signal/tmHMM) (Krogh et al. 2001; Petersen et al. 2011), and comparison to currently curated annotation databases (EMBL Uniprot eggNOG/GO Pathways databases) (Ashburner et al. 2000; Kanehisa et al. 2012; Powell et al. 2012).

Phylogenetic analysis: Predicted ORFs from the Trinity assembled transcripts were searched for selected genes and used for phylogenetic analyses. In these analyses, the selected protein sequences representing *G. janae* were aligned with the published sequences of apicomplexan members using MUSCLE (Version 3.8.31) sequence alignment software (Edgar 2004a, b). The resulting multiple sequence alignment was curated and manually edited, using BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html) to remove uninformative positions of the alignment. In parallel, Gblocks program (Version 0.91b) (Castresana 2000; Talavera and Castresana 2007) was used to curate the alignment with the highest stringency parameters. Maximum likelihood and an improved method of neighbor joining phylogenetic analyses were carried out using PhyML 3.1/3.0 aLRT (Guindon et al. 2010) and BioNJ (Gascuel 1997), respectively, to assess the evolutionary relationship between *G. janae* species and other members of the phylum Apicomplexa. The MUSCLE alignments, manually curated and Gblocks curated, were fed into PhyML for tree construction (SH-like approximate Likelihood-Ratio Test), using WAG model of amino acids substitution with NNI topology search (Fig. 3 and Supplementary Material Fig. S3 respectively). The tree generation in BioNJ was done by the James-Taylor-Thornon (JTT) matrix substitution model with 1000 bootstraps (Supplementary Material Fig. S4). The sequences representing Ciliophora (*Tetrahymena thermophila / Stylonychia lemnae*) were used as an outgroup to assess the position of *G. janae* with respect to other apicomplexans. The sequences used in the phylogenetic analysis were listed in Supplementary Material Table S6, and the alignment files before and after curation as Supplementary Material Figure S5. Phylogeny.fr platform was utilized for much of the above analysis (Dereeper et al. 2008, 2010).

Acknowledgements

This work was supported by the Czech Science Foundation (Centre of Excellence, P505/12/G112). Sunil Kumar Dogga was supported by the MalarX program, from the Swiss Initiative in Systems Biology (SystemsX.ch). RNA Sequencing on the Illumina HiSeq 2500 was performed at the iGE3 genomics platform of the University of Geneva (http://www.ige3.unige.ch/genomics-platform.php). The computations were performed at University of Geneva on the Baobab cluster. We thank Eva Dobaková for RNA isolation and Nicolas Hulo for his expert assistance on the analysis of the transcriptomic data. We thank Arnault Graindorge for his assistance on the phylogenetic analysis. We also thank Damien Jacot, Karine Frenal, Stepan Tymoshenko, Rebecca Oppenheim, George Rugarabamu and Hayley Bullen for their guidance and comments during the writing of the manuscript.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.protis.2015.09.003.

References

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, Buck GA, Xu P, Bankier AT, Dear PH, Konfortov BA, Spriggs HF, Iyer L, Anantharaman V, Aravind L, Kapur V (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. Science **304**:441–445

Alexander DL, Mital J, Ward GE, Bradley P, Boothroyd JC (2005) Identification of the moving junction complex of *Toxoplasma gondii*: A collaboration between distinct secretory organelles. PLoS Pathog **1**:e17

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215**:403–410

Anderson-White BR, Ivey FD, Cheng K, Szatanek T, Lorestani A, Beckers CJ, Ferguson DJ, Sahoo N, Gubbels MJ (2011) A family of intermediate filament-like proteins is sequentially assembled into the cytoskeleton of *Toxoplasma gondii*. Cell Microbiol **13**:18–31

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25**:25–29

Bairoch A, Boeckmann B, Ferro S, Gasteiger E (2004) Swiss-Prot: juggling between evolution and stability. Brief Bioinform **5**:39–55

Barta JR, Martin DS, Liberator PA, Dashkevicz M, Anderson JW, Feighner SD, Elbrecht A, Perkins-Barrow A, Jenkins MC, Danforth HD, Ruff MD, Profous-Juchelka H (1997) Phylogenetic relationships among eight eimeria species infecting domestic fowl inferred using complete small subunit ribosomal DNA sequences. J Parasitol **83**:262–271

Bartošová-Sojková P, Oppenheim RD, Soldati-Favre D, Lukeš J (2015) Epicellular apicomplexans: parasites "on-the-way-in". PLoS Pathog **11**:e1005080

Beck JR, Rodriguez-Fernandez IA, Cruz de Leon J, Huynh M-H, Carruthers VB, Morrissette NS, Bradley PJ (2010) A novel family of *Toxoplasma* IMC proteins displays a hierarchical organization and functions in coordinating parasite division. PLoS Pathog **6**:e1001094

Besteiro S, Michelin A, Poncet J, Dubremetz J-F, Lebrun M (2009) Export of a *Toxoplasma gondii* rhoptry neck protein complex at the host cell membrane to form the moving junction during invasion. PLoS Pathog **5**:e1000309

Boucher LE, Bosch J (2015) The apicomplexan glideosome and adhesins – Structures and function. J Struct Biol **190**:93–114

Bullen HE, Tonkin CJ, O'Donnell RA, Tham W-H, Papenfuss AT, Gould S, Cowman AF, Crabb BS, Gilson PR (2009) A novel family of apicomplexan glideosome-associated poteins with an inner membrane-anchoring role. J Biol Chem **284**:25353–25363

Carruthers V, Boothroyd JC (2007) Pulling together: an integrated model of *Toxoplasma* cell invasion. Curr Opin Microbiol **10**:83–89

Carruthers VB, Tomley FM (2008) Microneme proteins in apicomplexans. Subcell Biochem **47**:33–45

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol **17**:540–552

Chen AL KE, Toh JY, Vashisht AA, Rashoff AQ, Van C, Huang AS, Moon AS, Bell HN, Bentolila LA, Wohlschlegel JA, Bradley PJ (2015) Novel components of the *Toxoplasma* inner membrane complex revealed by BioID. mBio 6.

Chen F, Mackey AJ, Stoeckert CJ, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res **34**:D363–D368

Chen J, Zhou DH, Li ZY, Petersen E, Huang SY, Song HQ, Zhu XQ (2014) *Toxoplasma gondii*: protective immunity induced by rhoptry protein 9 (TgROP9) against acute toxoplasmosis. Exp Parasitol **139**:42–48

Consortium TU (2015) UniProt: a hub for protein information. Nucleic Acids Res **43**:D204–D212

Coppin A, Dzierszinski F, Legrand S, Mortuaire M, Ferguson D, Tomavo S (2003) Developmentally regulated biosynthesis of carbohydrate and storage polysaccharide during differentiation and tissue cyst formation in *Toxoplasma gondii*. Biochimie **85**:353–361

Danne JC, Gornik SG, Macrae JI, McConville MJ, Waller RF (2013) Alveolate mitochondrial metabolic evolution: dinoflagellates force reassessment of the role of parasitism as a driver of change in apicomplexans. Mol Biol Evol **30**:123–139

Dereeper A, Audic S, Claverie JM, Blanc G (2010) BLAST-EXPLORER helps you building datasets for phylogenetic analysis. BMC Evol Biol **10**:8

Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O (2008) Phylogeny. fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res **36**:W465–W469

Dyková I, Lom J (1981) Fish coccidia: critical notes on life cycles, classification and pathogenicity. J Fish Dis **4**:487–505

Edgar RC (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**:113

Edgar RC (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32**:1792–1797

Entrala E, Mascaró C (1997) Glycolytic enzyme activities in *Cryptosporidium parvum* oocysts. FEMS Microbiol Lett **151**:51–57

Ferguson DJ, Parmley SF, Tomavo S (2002) Evidence for nuclear localisation of two stage-specific isoenzymes of enolase in *Toxoplasma gondii* correlates with active parasite replication. Int J Parasitol **32**:1399–1410

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res **39**:W29–W37

Fleige T, Limenitakis J, Soldati-Favre D (2010) Apicoplast: keep it or leave it. Microbes Infect **12**:253–262

Fournie JW, Overstreet RM (1983) True intermediate hosts for *Eimeria funduli* (Apicomplexa) from estuarine fishes. J Protozool **30**:672–675

Frénal K, Marq J-B, Jacot D, Polonais V, Soldati-Favre D (2014) Plasticity between MyoC- and MyoA-glideosomes: An example of functional compensation in *Toxoplasma gondii* invasion. PLoS Pathog **10**:e1004504

Frénal K, Polonais V, Marq J-B, Stratmann R, Limenitakis J, Soldati-Favre D (2010) Functional dissection of the apicomplexan glideosome molecular architecture. Cell Host Microbe **8**:343–357

Fritz HM, Buchholz KR, Chen X, Durbin-Johnson B, Rocke DM, Conrad PA, Boothroyd JC (2012) Transcriptomic analysis of *Toxoplasma* development reveals many novel functions

and structures specific to sporozoites and oocysts. PLoS ONE **7**:e29998

Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, Heiges M, Iodice J, Kissinger JC, Mackey AJ, Pinney DF, Roos DS, Stoeckert CJ, Wang H, Brunk BP (2008) ToxoDB: an integrated *Toxoplasma gondii* database resource. Nucleic Acids Res **36**:D553–D556

Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol **14**:685–695

Gaskins E, Gilk S, DeVore N, Mann T, Ward G, Beckers C (2004) Identification of the membrane receptor of a class XIV myosin in *Toxoplasma gondii*. J Cell Biol **165**:383–393

Gold DA, Kaplan AD, Lis A, Bett GC, Rosowski EE, Cirelli KM, Bougdour A, Sidik SM, Beck JR, Lourido S, Egea PF, Bradley PJ, Hakimi MA, Rasmusson RL, Saeij JP (2015) The *Toxoplasma* dense granule proteins GRA17 and GRA23 mediate the movement of small molecules between the host and the parasitophorous vacuole. Cell Host Microbe **17**: 642–652

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech **29**:644–652

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3. 0. Syst Biol **59**:307–321

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc **8**: 1494–1512

Harding CR, Meissner M (2014) The inner membrane complex through development of *Toxoplasma gondii* and *Plasmodium*. Cell Microbiol **16**:632–641

Henschel R, Lieber M, Wu L-S, Nista PM, Haas BJ, LeDuc RD (2012) Trinity RNA-Seq Assembler Performance Optimization. In Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond. ACM, Chicago, Illinois, USA, pp 1–8

Jirků M, Jirků M, Oborník M, Lukeš J, Modrý D (2009) *Goussia* Labbé, 1896 (Apicomplexa, Eimeriorina) in Amphibia: Diversity, biology, molecular phylogeny and comments on the status of the genus. Protist **160**:123–136

Jirků M, Modrý D, Šlapeta JR, Koudela B, Lukeš J (2002) The phylogeny of *Goussia* and *Choleoeimeria* (Apicomplexa; Eimeriorina) and the evolution of excystation structures in coccidia. Protist **153**:379–390

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res **40**:D109–D114

Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, Browne P, van den Broek A, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Gamble J, Diez FG, Harte N, Kulikova T, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Sobhany S, Stoehr P, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R (2005) The EMBL Nucleotide Sequence Database. Nucleic Acids Res **33**:D29–D33

Katris NJ, van Dooren GG, McMillan PJ, Hanssen E, Tilley L, Waller RF (2014) The apical complex provides a regulated gateway for secretion of invasion factors in *Toxoplasma*. PLoS Pathog **10**:e1004074

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1. J Mol Biol **305**:567–580

Kuo C-H, Wares JP, Kissinger JC (2008) The apicomplexan whole-genome phylogeny: an analysis of incongruence among gene trees. Mol Biol Evol **25**:2689–2698

Lamarque MH, Roques M, Kong-Hap M, Tonkin ML, Rugarabamu G, Marq J-B, Penarete-Vargas DM, Boulanger MJ, Soldati-Favre D, Lebrun M (2014) Plasticity and redundancy among AMA–RON pairs ensure host cell entry of *Toxoplasma* parasites. Nat Commun **5**:4098

Limenitakis J, Oppenheim RD, Creek DJ, Foth BJ, Barrett MP, Soldati-Favre D (2013) The 2-methylcitrate cycle is implicated in the detoxification of propionate in *Toxoplasma gondii*. Mol Microbiol **87**:894–908

Lom J, Dyková J (1992) Protozoan Parasites of Fishes. Elsevier Sci Publ, Amsterdam, 315 p

Lovy J, Friend SE (2015) Intestinal coccidiosis of anadromous and landlocked alewives, *Alosa pseudoharengus*, caused by *Goussia ameliae* n. sp. and *G. alosii* n. sp. (Apicomplexa: Eimeriidae). Int J Parasitol Parasites Wildl **4**: 159–170

Lukeš J, Dyková I (1990) *Goussia janae* n. sp. (Apicomplexa, Eimeriorina) in dace *Leuciscus leuciscus* and chub *L. cephalus*. Dis Aquat Organ **8**:85–90

Lukeš J, Starý V (1992) Ultrastructure of the life-cycle stages of *Goussia janae* (Apicomplexa. Eimeriidae), with X-ray microanalysis of accompanying precipitates. Can J Zool **70**:2382–2397

Maslov DA, Lukeš J, Jirků M, Simpson L (1996) Phylogeny of trypanosomes as inferred from the small and large subunit rRNAs: Implications for the evolution of parasitism in the trypanosomatid protozoa. Mol Biochem Parasitol **75**: 197–205

Mazurie AJ, Alves JM, Ozaki LS, Zhou S, Schwartz DC, Buck GA (2013) Comparative genomics of *Cryptosporidium*. Int J Genomics **2013**:832756

Molnár K (1996) Remarks on the morphology, site of infection and validity of some coccidian species from fish. Acta Vet Hung **44**:295–307

Molnár K, Ostoros G, Dunams-Morel D, Rosenthal BM (2012) *Eimeria* that infect fish are diverse and are related to, but distinct from, those that infect terrestrial vertebrates. Infect Genet Evol **12**:1810–1815

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

Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4. 0: discriminating signal peptides from transmembrane regions. Nat Meth **8**:785–786

Poukchanski A, Fritz HM, Tonkin ML, Treeck M, Boulanger MJ, Boothroyd JC (2013) *Toxoplasma gondii* sporozoites invade host cells using two novel paralogues of RON2 and AMA1. PLoS ONE **8**:e70637

Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, Jensen LJ, von Mering C, Bork P (2012) eggNOG v3. 0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res **40**:D284–D289

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD (2012) The Pfam protein families database. Nucleic Acids Res **40**:D290–D301

Saito T, Nishi M, Lim MI, Wu B, Maeda T, Hashimoto H, Takeuchi T, Roos DS, Asai T (2008) A novel GDP-dependent pyruvate kinase isozyme from *Toxoplasma gondii* localizes to both the apicoplast and the mitochondrion. J Biol Chem **283**:14041–14052

Seeber F, Limenitakis J, Soldati-Favre D (2008) Apicomplexan mitochondrial metabolism: a story of gains, losses and retentions. Trends Parasitol **24**:468–478

Shanmugasundram A, Gonzalez-Galarza FF, Wastling JM, Vasieva O, Jones AR (2013) Library of apicomplexan metabolic pathways: a manually curated database for metabolic pathways of apicomplexan parasites. Nucleic Acids Res **41**:D706–D713

Sivagurunathan S, Heaslip A, Liu J, Hu K (2013) Identification of functional modules of AKMT, a novel lysine methyltransferase regulating the motility of *Toxoplasma gondii*. Mol Biochem Parasitol **189**:43–53

Steinhagen D, Körting W (1990) The role of tubificid oligochaetes in the transmission of *Goussia carpelli*. J Parasitol **76**:104–107

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol **56**:564–577

Tyler JS, Treeck M, Boothroyd JC (2011) Focus on the ringleader: the role of AMA1 in apicomplexan invasion and replication. Trends Parasitol **27**:410–420

Van Bel M, Proost S, Van Neste C, Deforce D, Van de Peer Y, Vandepoele K (2013) TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. Genome Biol **14**, R134–R134

van Dooren GG, Striepen B (2013) The algal past and parasite present of the apicoplast. Annu Rev Microbiol **67**:271–289

Whipps CM, Fournie JW, Morrison DA, Azevedo C, Matos E, Thebo P, Kent ML (2012) Phylogeny of fish-infecting *Calyptospora* species (Apicomplexa: Eimeriorina). Parasitol Res **111**:1331–1342

Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, Pearson WR, Dear PH, Bankier AT, Peterson DL, Abrahamsen MS, Kapur V, Tzipori S, Buck GA (2004) The genome of *Cryptosporidium hominis*. Nature **431**:1107–1112

Yang S, Parmley SF (1997) *Toxoplasma gondii* expresses two distinct lactate dehydrogenase homologous genes during its life cycle in intermediate hosts. Gene **184**:1–12