Available online at www.sciencedirect.com
SCIENCE DIRECT
International Journal for Parasitology 34 (2004) 1265–1271
www.parasitology-online.com

Expressed sequence tag (EST) analysis of Gregarine gametocyst development☆

Charlotte K. Omoto[a,*], Marc Toso[a], Keliang Tang[b], L. David Sibley[b]

[a]School of Biological Sciences, Washington State University, Pullman, WA 99164-4236, USA
[b]Department of Molecular Microbiology, Washington University School of Medicine, St Louis, MO 63110, USA

Received 13 July 2004; received in revised form 16 August 2004; accepted 17 August 2004

Abstract

Gregarines are protozoan parasites of invertebrates in the phylum Apicomplexa. We employed an expressed sequence tag strategy in order to dissect the molecular processes of sexual or gametocyst development of gregarines. Expressed sequence tags provide a rapid way to identify genes, particularly in organisms for which we have very little molecular information. Analysis of ~1800 expressed sequence tags from the gametocyst stage revealed highly expressed genes related to cell division and differentiation. Evidence was found for the role of degradation and recycling in gametocyst development. Numerous additional genes uncovered by expressed sequence tag sequencing should provide valuable tools to investigate gametocyst development as well as for molecular phylogenetics, and comparative genomics in this important group of parasites.

Keywords: Apicomplexa; Gene expression; Cell wall; Histones

1. Introduction

Gregarines are protozoa of the phylum Apicomplexa, which is comprised of obligate parasites. A few members of this phylum are intensively studied because they have great medical and veterinary impact such as *Plasmodium*, *Toxoplasma*, *Eimeria*, *Neospora* and *Cryptosporidium*. The complete genomes of *Plasmodium falciparum* (Gardner et al., 2002), and *Cryptosporidium parvum* (Abrahamsen et al., 2004) have been sequenced and other genome projects of apicomplexans are in progress (http://www.sanger.ac.uk/Projects/Protozoa/, http://www.tigr.org/tdb/parasites/). In contrast, gregarines are understudied. Three subdivisions of gregarines are recognised based on morphology and life cycle characteristics. Archigregarines, as the name implies, are presumed to be the most primitive and parasitize annelids, eugregarines parasitize a variety of invertebrates including insects, and neogregarines parasitize insects.

Gregarines have a remarkable life cycle that is completed in a single invertebrate host (Fig. 1). The life cycle of *Gregarina niphandrodes* begins with the infectious haploid sporozoites being ingested by the host. The sporozoites transform into trophozoites (Fig. 1A), the feeding stage, and increase in size. The sexual stage commences with two trophozoites joining in a stage called syzygy (Fig. 1B). The trophozoites at this stage are referred to as gamonts, and they encyst and become a gametocyst (Fig. 1C). The gametocysts are released within the host feces. Dramatic cell differentiation and multiplication occurs inside the gametocyst after shedding from the host. First, the haploid gamonts divide into hundreds of male and female gametes. Using their flagella, the male gametes mix the cells and facilitate fertilisation. The fertilised zygote, sometimes called an oocyst then undergoes meiosis and mitosis to form eight haploid spores or sporozoites within sporocysts that are released in chains (Fig. 1D) to begin the cycle anew.

☆Sequences described in this paper are deposited in GenBank under accession numbers CF275644–CF276012, CF945836–CF947046, CK401114–CK401260, CO635960-CO636151.

* Corresponding author. Tel.: +1 509 335 5591; fax: +1 509 335 3184.
E-mail address: omoto@wsu.edu (C.K. Omoto).

0020-7519/$30.00
doi:10.1016/j.ijpara.2004.08.004

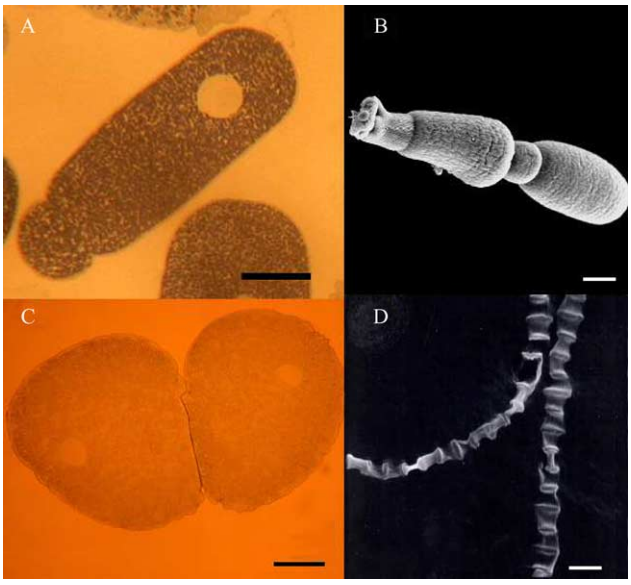

Fig. 1. Micrographs of four stages in the life cycle of *Gregarina niphandrodes*. (A) Light micrograph of a section of trophozoite. Scale bar 500 μm. (B) Scanning electron micrograph of two trophozoites in syzygy. Scale bar 352 μm. (C) Light micrograph of a section of encysting gamonts. Scale bar 500 μm. (D) Scanning electron micrograph of chains of sporocysts. Scale bar 60 μm.

Although the general outline of gregarine gametocyst development has been appreciated for decades, the underlying cellular and molecular processes are not known.

As of the end of 2003, there were only 22 gregarine DNA sequences deposited in GenBank, the majority of which are of the small subunit ribosomal RNA gene. In addition, sequences for a few cytoskeletal protein-encoding genes from gregarines (i.e. actins, myosins, and tubulins) are available in GenBank. Consequently, we employed an EST strategy to investigate the sexual stage of the life cycle and to increase the rate of gene discovery in gregarines.

As a first step in the molecular dissection of gametocyst development, we have analysed a collection of EST sequences from randomly chosen cDNAs of gametocyst stage. Expressed sequence tags (Boguski et al., 1993) provide a rapid way to identify genes, particularly in organisms for which we have very little molecular information. A tremendous increase in cell number concomitant with dramatic cell differentiation occurs in gametocysts. Consequently, EST analysis should prove to be particularly informative in identifying genes involved in gametocyst development including cell differentiation, cell cycle regulation, and cell division in gregarines.

Obtaining quantities of clean material from gregarines is a challenge since methods for growing them *in vitro* have not been developed. Yet, *G. niphandrodes* provides a good species for this initial attempt. *Gregarina niphandrodes* can be propagated in their host *Tenebrio molitor*, mealworm beetles. Gametocysts of *G. niphandrodes* are unusually tough; thus they can be extensively cleaned of contaminating host and fecal material. The ability to obtain quantities of these gametocysts allowed us to generate the first cDNA library and commence sequencing of gregarine ESTs.

2. Materials and methods

2.1. Collection of G. niphandrodes gametocysts

Tenebrio molitor were placed in 150 mm Petri dish for ~24 h to collect gametocysts of *G. niphandrodes* from the shed fecal material. The material was suspended in water and agitated in sterile distilled water to separate them from the fecal material, and allowed to settle under $1 \times g$ gravity in a step sucrose gradient. The white cysts are visible to the naked eye and banded between the 10 and 20% sucrose steps. They were collected manually with a Pasteur pipette and extensively washed in sterile distilled water, then in diethylpyrocarbonate (DEPC)-treated water. They were then placed in a minimum volume of Trizol (Invitrogen) sufficient to cover them and stored at −70°. Because *T. molitor* was left in the Petri dish for ~24 h, gametocysts range in age from 0, freshly shed, to 24 h, by which time the sporocysts are almost ready to dehisce (unpublished observations). Thus, we expect all stages of gametocyst development to be represented in the collected cysts.

2.2. RNA extraction, cDNA library construction and EST sequencing

Total RNA was extracted from washed gametocysts by digestion in Trizol (Invitrogen), phenol–chloroform extraction, and precipitation in ethanol. Messenger RNA was isolated from total RNA using oligotex mRNA spin columns (Qiagen). cDNA was synthesised from poly(A)+ mRNA using the template-switching PCR method (SMART cDNA Kit, BD Biosciences). First strand cDNA was reverse transcribed using the CDS III/3′ primer and a 5′ template switch primer (Smart IV primer). The product of the first strand synthesis was PCR amplified using the same primer set and the fragments were digested with SfiI. The fragments were size selected, ligated into a modified pBluescript vector (obtained from Michael White, Montana State University) containing directional SfiI sites, and electroporated into ElectroTen Blue *Escherichia coli* cells. The vector was modified in the following way: SfiI sites were added to the multiple cloning region of pBluescript SK+ between the BamHI/EcoRI sites. The modified polylinker has the following sequence: 5′GAATTCGGCCATTACGGCC (G)*n*–insert–GGCCGCCTCGGCCCACGGATCC3′ (where $n=3-4$). Individual clones were selected robotically and subjected to large-scale sequencing as described previously (Li et al., 2003). Sequences were derived from the 5′ and 3′ ends independently using universal primers that were complementary to the vector. In brief, sequences were generated

using BigDye terminator chemistry (ABI) and reactions were resolved on ABI 3700 capillary sequences. DNA traces were evaluated using Gelminder and EST_OTTO and then processed to generate high quality basecalls and trim vector sequences as described previously (Li et al., 2003). The great majority of DNA sequencing was performed by Washington University Genome Sequencing Center and 192 sequences by Amplicon (Pullman, WA).

2.3. Analysis of sequences

Sequences were initially compared using BLASTX searches against the SWIR (release 21), a non-redundant database consisting of sequences from PIR, SWIS-PROT and WORMPEP. The homologies identified are not due to low complexity since low complexity sequences were filtered in the BLAST searches. These results were annotated at the time of submission to the dbEST section of NCBI (http://www.ncbi.nlm.nih.gov/dbEST/index.html).

EST sequences were processed and annotated using PipeOnline (POL) (Ayoubi et al., 2002) (http://bioinfo.okstate.edu/pipeonline/). This database is searchable by the description of the closest homologs, organism name of the closest homologs, as well as by clone id# and contig id#.

Individual cDNAs were amplified by 3′ RACE and sequenced to determine the 3′ untranscribed region and polyadenylation site. Internal gene-specific primers (sequence available upon request) were used in combination with vector primers from the SMART cDNA synthesis kit to amplify the 3′ end of the bacterial-like genes (Table 2). Contigs were reassembled in Vectro NTi. Dinucleotide relative abundance was calculated using a Perl program written by one of the authors (C.K.O.), and is available upon request. Mitoprot (http://ihg.gsf.de/ihg/mitoprot.html) (Claros and Vincens, 1996) was used to calculate the probability of a mitochondrial targeting sequence appearing in a contig.

2.4. Microscopy

Gregarina niphandrodes trophozoites and gamonts were collected from the intestine of adult *T. molitor* and fixed in 2% paraformaldehyde and 2% glutaraldehyde in 0.1 M cacodylate buffer overnight. Samples were rinsed three times for 10 min each in 0.1 M cacodylate buffer and post-fixed in 2% osmium tetroxide for 1 h. They were then washed in 0.1 M cacodylate buffer three times for 10 min each. The specimens were dehydrated in a graded acetone series and embedded in Spurr's resin. For light microscopy, 1 μm thick sections were cut on a Leica Reichert microtome, mounted on gelatin-coated slides and stained with 0.1% toluidine blue in water at 60 °C for 2 min.

For scanning electron micrographs of the spore chains, gametocysts were allowed to dehisce under humid atmosphere for 24–48 h, placed on stubs, and coated with gold. Trophozoites in sygyzy were dehydrated in a graded series of acetone and hexamethydisilazane to 100% hexamethydisilazane. They were air dried in a vial overnight, mounted on stubs and coated with gold. Both samples were viewed with Hitachi S570 scanning electron microscope.

3. Results and discussion

A total of 1919 ESTs for *G. niphandrodes* were submitted to NCBI. PipeOnline (http://bioinfo.okstate.edu/pipeonline/) was used to cluster the ESTs into contigs and assign them into functional categories. The PipeOnline database is searchable by the name of the protein or the name of the organism with the closest homolog. The data in PipeOnline can also be browsed to examine contigs with high scoring pairs, expectation, or bit-score criteria. We chose to consider only those contigs that had matches to identified genes in the database with E values less than 1×10^{-4}. We also included unmatched contigs longer than 300 bases. This pruning left 1854 ESTs that clustered into 481 contigs. Of these, 725 ESTs in 147 contigs had significant matches (E value $<1\times10^{-4}$) to known proteins in the database. An additional 215 ESTs in 30 contigs had significant matches (E value $<1\times10^{-4}$) to hypothetical proteins in the database. This left 914 ESTs in 304 contigs with no significant match to proteins in the database.

During gametocyst development, there is ~1000-fold increase in cell number involving four major developmental processes: gametogenesis, fertilisation, meiosis, and spore formation. Thus, many of the identified ESTs are from genes that are expected to be involved in these processes (Table 1). Contigs that were most highly represented in the EST database were histones with 360 ESTs or over 19% of the EST analysed with 154 ESTs corresponding to histone H2b. Over 100 ESTs formed a contig with the closest match to histone 60, the germinal histone H4 in *Caenorhabditis elegans* (Maeda et al., 2001). Oocyst wall protein, another structural protein expected in gametocyst development, is represented by 76 ESTs with highly significant matches to the cysteine-rich repeat oocyst wall protein with from *C. parvum*. This finding suggests that the basic cell wall architecture between the oocyst in coccidian parasites and the gametocysts in gregarines is highly conserved. Tubulins are needed throughout gametocyst development for mitotic and meiotic spindles as well as for the flagella, and are represented by 18 ESTs. The β-tubulin ESTs have a highly significant match to the gregarine β-tubulin sequence in the database. In addition, ESTs with a high match to a dynein heavy chain and a dynein light chain from *Cryptosporidium* were identified (Table 1).

We considered genes in the information pathway to include DNA replication, DNA repair, transcription, and translation. ESTs with significant matches to proteins in the pathway were represented in the database with a total of 41 contigs consisting of 87 ESTs. Among these are contigs with high matches to DNA polymerase, DNA helicase,

Table 1
Putative identification of genes expressed during gametocyst development

Gene name or functional class	# of ESTs	Best match[a]		Other matches of interest[b]		
		E value	Genus	*E* value	Genus	GI #[c]
Histones						
H2a	53	2×10^{-36}	*Schizosaccharomyces*			19115333
H2b	154	7×10^{-24}	*Tetraodon*	6×10^{-22}	*Plasmodium*	23508258
H3	35	1×10^{-49}	*Dermasterias*	4×10^{-27}	*Plasmodium*	23619278
H4	118	6×10^{-43}	*Caenorhabditis*			17540668
Oocyst wall protein	76	5×10^{-51}	*Cryptosporidium*			32398788
Tubulin						
α-Tubulin	11	4×10^{-55}	*Cryptosporidium*			21634435
β-Tubulin	7	1×10^{-173}	*Leidyana* (gregarine)			25991883
Dynein						
dynein heavy chain, cytoplasmic	1	1×10^{-13}	*Tetrahymena*			13561925
dynein light chain	1	2×10^{-14}	*Chlamydomonas*			7484373
Genes in the information pathway						
DNA helicase	11	9×10^{-16}	*Plasmodium*			18026950
DNA polymerase	10	3×10^{-51}	*Cryptosporidium*			323990
DNA topoisomerase	2	2×10^{-31}	*Plasmodium*			23491424
RecA	6	6×10^{-59}	*Cryptosporidium*			46227223
Elongation factor	12	1×10^{-82}	*Cryptosporidium*			1737177
Translation initiation factor	5	5×10^{-34}	*Plasmodium*			2349637
DNA repair[d]	8					
Ribosomal proteins[d]	16					
Others[d]	17					
Genes involved in cell cycle regulation and cell signaling						
Protein kinases and regulatory subunit[d]	26	2×10^{-56}	*Toxoplasma*			
Calmodulin	5	2×10^{-58}	*Stylonichia*			161195
cAMP phosphodiesterase	3	1×10^{-39}	*Plasmodium*			23619164
Protein phosphatase[d]	4					
Others[d]	2					
Autophagy and vesicular trafficking						
Autophagy	10	1×10^{-38}	*Plasmodium*			23480688
Clathrin heavy chain	4	1×10^{-11}	*Plasmodium*			23488159
Proteasome	8	4×10^{-49}	*Plasmodium*			23479087
Ubiquitin[d]	7	1×10^{-89}	*Drosophila*	7×10^{-24}	*Plasmodium*	2361308
Ubiquitin conjugating enzyme[d]	8	2×10^{-48}	*Plasmodium*			23619484
Chitinase	3	8×10^{-11}	*Coxiella*			29655324
Vacuolar import and degradation	3	5×10^{-32}	*Cryptosporidium*			46226914
Others	6					
Transport proteins						
ATP binding cassette transporter	6	7×10^{-37}	*Drosophila*			24643674
Sugar transport proteins[d]	2	5×10^{-24}	*Cryptosporidium*			46228482

[a] The best BLASTx match *E* value is shown only for specific proteins. The best BLASTx match genus is shown only for specific proteins or when the group of proteins identified the same species as the highest match.
[b] When there are two taxonomic group matches listed, the gi number refers to the other match of interest.
[c] GI refers to the gene index which can be used to search through GenBank.
[d] Indicates that a number of different recognised genes are grouped together.

elongation factor, and ribosomal proteins (Table 1). We also found contigs with high matches to genes required for cell cycle regulation and cell signaling. These include protein kinases and its regulatory subunit and protein phosphatases (Table 1).

There are some genes not identified among our EST database that we expected to observe. One such gene is actin. Actins are relatively conserved proteins, and sequences from two gregarine species, including one complete sequence, are in GenBank. Gregarine trophozoites

have an extensive actin filament network, and actin has been proposed as necessary for their unique gliding motility (Chen and Fan-Chiang, 2001; Schrevel and Philippe, 1993; Heintzelman, 2004). Previous EST projects on apicomplexans have found that actin and actin-depolymerising factor genes are among the most abundantly transcribed genes of motile stages (Li et al., 2003). The failure to detect these genes here suggests that cells in the gametocyst stage do not undergo gliding motility.

We were surprised to find contigs with significant matches to proteins involved in degradation (Table 1). Ubiquitin plays an essential role in targeting proteins for proteasome-mediated degradation. Contigs with highly significant matches to proteasome subunits (eight ESTs), and ubiquitin and ubiquitin-conjugating enzymes were found (15 ESTs). Autophagy is a process for recycling of macromolecules that occurs via engulfment of organelles and lysosomal degradation (Wang and Klionsky, 2003). A contig of 10 ESTs has a significant match to an autophagy protein from *Plasmodium yoelii*. Disruption of the homologous gene in *Arabidopsis* (GI#19912167) accelerates leaf senescence (Hanaoka et al., 2002). In addition, there are contigs with significant matches to vacuolar import and degradation protein vid27, and clathrin heavy chain. We also found three ESTs with matches to chitinase in the diverse chitinase family 18, which contains chitinases from bacteria, animals, and plants (Henrissat and Davies, 2000). Chitinase has been implicated in *Plasmodium* ookinete invasion of the peritrophic membrane of the mosquito (Langer and Vinetz, 2001). However, since our ESTs are from the stage that occurs outside the host, it may mean that chitin is a structural component in gregarines and their breakdown is a necessary step in gametocyst development. Chitinase has been implicated in sporogonic development of *Plasmodium* (Bhatnagar et al., 2003). Alternatively, these proteins may be involved in breakdown of storage polysaccharides. The identification of the spectrum of genes involved in degradation and recycling of components suggests that these processes play an important role in gametocyst development.

We identified ESTs with a significant match to ATP-binding cassette transporter and sugar transport (Table 1). We expect to find genes involved in transport to be expressed in the trophozoite stage since they may be necessary for transport of nutrients from the host intestinal lumen. However, we did not expect these genes to be expressed at a significant level in the gametocyst stage that is outside the host and totally enclosed in a tough cyst wall.

There are some highly represented genes with matches only to hypothetical proteins. For example, one contig of 124 ESTs consisting of 1839 bp has a significant match (4×10^{-18}) to a hypothetical protein in *P. yoelii* genome (GI#23488485). There are 11 additional contigs with highly significant ($<<10^{-20}$) matches to hypothetical proteins from *Plasmodium* or *Cryptosporidium*. Thus we have identified genes that are highly expressed during gregarine gametocyst development that also occur in other Apicomplexa. We also found evidence for gregarine-specific genes: one contig of 149 ESTs consisting of 639 bp has no match to the database, despite being abundant in *G. niphandrodes*. Such genes may present unique pathways in gregarine development.

Surprisingly, we found 28 ESTs with highly significant ($<<10^{-50}$) matches to bacterial genes (Table 2). In particular, one 1279 bp contig consisting of nine ESTs had a highly significant match to branched-chain amino acid aminotransferase from *Bacteroides* and no significant match to any apicomplexan. This contig contained a 14 bp polyA tail suggesting it is a bone fide gregarine gene and not due to bacterial contamination. In contrast, another contig of seven ESTs had the highest match to ornithine aminotransferase from *Cytophaga* but also had a highly significant match to the gene from *Plasmodium*. A third contig with seven ESTs had the highest match to mannitol-1-phosphate dehydrogenase from *Shigella* and no significant match to any apicomplexan. We performed 3′ RACE reactions based on the consensus sequence for these two EST contigs and found

Table 2
ESTs with highly significant match ($<1\times10^{-50}$) to bacterial genes

Gene name	#ESTs	polyA?	%GC	*E* value	Highest matches				
					Bacterial	GI#[a]	*E* value	Non-bacterial	GI#
Amino acid aminotransferase[a]	9	Yes	51	4×10^{-87}	*Bacteroides*	29349300	2×10^{-77}	*Giardia*	27980319
Mannitol-1-phosphate DH[b]	7	Yes	49	1×10^{-58}	*Shigella*	24114864	4×10^{-58}	*Dunaliella*	21483850
Ornithine aminotransferase[c]	7	Yes	53	5×10^{-56}	*Cytophaga*	23137751	1×10^{-49}	*Plasmodium*	23612170
Malate/lactate DH[d]	1	nd	55	4×10^{-53}	*Rhodospirillum*	48764981	1×10^{-42}	*Plasmodium*	23612264

GI# of ESTs: a. 33653282, 33653064, 33653100, 33653163, 33653219, 33653397, 38450919, 38451281, 38451333. b. 38451015, 38451049, 38451648, 38451733, 38451742, 40559967, 40560018. c. 38450678, 38451038, 38451448, 38451551, 38451757, 40559945, 50539343. d. 38450714. nd, not determined.

[a] GI refers to gene index which can be used to search through GenBank.

Table 3
Comparison of dinucleotide frequencies

Sequence	bps	%GC	Dinucleotides[a]	
			GC	TA
Gregarine[b]	10,420	49	1.00	0.66*
*EST (contigs shown in*Table 2)				
Amino acid transferase	1269	51	1.03	0.76*
Mannitol-1-phosphate DH	1342	49	1.07	0.74*
Ornithine aminotransferase	657	53	0.83	0.57*
Malate/lactate DH	591	55	0.95	0.54*
Bacterial genome[c]				
Bacteroides	21,021	45	1.06	0.80
Shigella	21,333	51	1.30*	0.68*
Cytophaga	21,234	38	1.24*	0.81

[a] Dinucleotide relative abundances that deviated from 1 are shown. *Indicates significant deviation from 1 ($0.78<$ or >1.23).
[b] Gregarine sequences were combined from different species. Analysis of individual genes all showed significant under representation of the dinucleotide TA.
[c] Bacterial sequences include the sequence of the highest match plus two different random 10,000 pieces of the sequenced genome.

that both contained poly A additions at the 3′ end of the transcripts (Table 2).

There are several possible explanations for contigs with highly significant matches to bacterial genes. One is that despite our attempt to clean the gametocysts of fecal material, they may have contained contaminating bacteria. Another possibility is that our gregarines contain intracellular bacterial endosymbionts. Bacteria-like structures in gregarines have been observed in thin section electron micrographs, including structures reminiscent of dividing bacteria (Mackenzie and Walker, 1979). However, the presence of a polyA tail on three of these contigs is more consistent with a process of horizontal gene transfer (Boucher and Doolittle, 2000; Huang et al., 2004). Finally, mitochondrial targeted genes may have significant homology to bacterial genes. We used Mitoprot to assess the probability that these sequences were targeted to the mitochondria. The contigs with highly significant match to bacterial sequences had very low or zero probability of being targeting to the mitochondria according to this algorithm.

Dinucleotide relative abundance provides one measure of the relatedness of DNA sequences (Campbell et al., 1999). Dinucleotide relative abundance is the frequency of dinucleotides compared to their expected frequency taking into account the GC content. The pattern of dinucleotide relative abundance is characteristic of different genomes and thus can be considered a ‘genomic signature’ (Campbell et al., 1999). Given a sufficiently long random sequence (>5 kb), dinucleotide relative abundance ranges outside <0.78 or >1.23, occurs with probability <0.001 independent of GC content (Karlin and Cardon, 1994). Thus we used this as the threshold values to indicate significant deviation from one, the expected frequency. We used all currently available protein encoding genes from gregarine species in GenBank (RPB1 (GI# 37726868), actin (GI# 37105911), β-tubulin (GI# 25991883), myosins (GI#s 37105935, 37105923)) with a total of ∼10 kb of sequence (49% GC) to calculate the gregarine dinucleotide relative abundance (Table 3). Only the dinucleotide TA (0.66) was significantly different from expected. Four contigs with high matches to bacterial sequences also had only an under representation of the dinucleotide TA (0.57, 0.74, 0.76, and 0.54). Thus, their dinucleotide relative abundance is similar to that of gregarine sequences. The corresponding closely matched bacterial genomes (as represented by >20 kb of sequences each) also show under-representation of the dinucleotide TA, but only the *Shigella* genome shows deviation below the threshold value of 0.78. In contrast, both *Cytophaga* and *Shigella* genomes have an over representation of the dinucleotide GC that is not found in gregarines, or the four contigs with match to bacterial genes. Collectively, the evidence from dinucleotide relative abundance analysis is consistent with the ESTs being of gregarine, rather than bacterial origin.

In addition to elucidating the genes involved in gametocyst development, this EST project represents the first major attempt to identify protein-coding genes in gregarines. The tremendous amount of sequence data available for apicomplexans facilitated assigning putative identification to many of our ESTs. Gregarines are considered to be the basal lineage among the Apicomplexa (Barta, 1997). Several genes identified in our EST database will be useful for elucidating taxonomic relationships among the Apicomplexa and their relationship to other protists. These genes include DNA-dependent RNA polymerase II largest subunit, tubulins, elongation factor, and heat shock protein (Arisue et al., 2002; Dacks et al., 2002; Hirt et al., 1999; Keeling and Doolittle, 1996; Moriya et al., 2001; Stiller et al., 1998).

Our study represents the first attempt to systematically explore the gene diversity in gregarines, ancient members of the phylum Apicomplexa. Gregarines are most closely related to *Cryptosporidium* (Barta, 1997) but are also closely aligned to *Plasmodium* spp. in some phylogenetic reconstructions (Li et al., 2003). The preliminary analysis of genes abundantly transcribed in gametocysts should provide an impetus for more comprehensive attempts to define the transcriptome of gregarines and to sequence complete genomes from key members of this group. Aside from their intrinsic biological interest, such studies would be of tremendous value for comparison to apicomplexans that are pathogenic to animals and humans.

Acknowledgements

We thank John Janovy for teaching us how to culture *G. niphandrodes* in *T. molitor*, Michelle Martin for patiently collecting thousands of cysts, Martin Morgan with help with

Perl programming, and Andris Kleinhofs for critical reading of an early version of the manuscript. We are also grateful to Robert Cole, Sandy Clifton, Deana Pape and the Washington University Genome Sequencing Center for assistance with EST sequencing.

References

Abrahamsen, M.S., Templeton, T.J., Enomoto, S., Abrahante, J.E., Zhu, G., Lancto, C.A., Deng, M., Liu, C., Widmer, G., Tzipor, S., Buck, G.A., Xu, P., Bankier, A.T., Dear, P.H., Konfortov, B.A., Spriggs, H.F., Iyer, L., Anantharaman, V., Aravind, L., Kapur, V., 2004. Complete genome sequence of the apicomplexan *Cryptosporidium parvum*. Science 304, 441–445.

Arisue, N., Hashimoto, T., Lee, J., Moore, D., Gordon, P, Sensen, C., Gaasterland, T., Hasegawa, M., Muller, M., 2002. The phylogenetic position of the pelobiont Mastigamoeba balamuthi based on sequences of rDNA and translation elongation factors EF1 alpha and EF-2. J. Eukaryot. Microbiol. 49, 1–10.

Ayoubi, P., Jin, X., Leite, S., Liu, X., Martajaja, J., Abduraham, A., Wan, Q., Yan, W., Misawa, E., Prade, R.A., 2002. PipeOnline 2.0, automated EST processing and functional data sorting. Nucleic Acid Res. 30, 4761–4769.

Barta, J.R., 1997. Investigating phylogenetic relationships within the Apicomplexa using sequence data: the search for homology. Methods 13, 81–88.

Bhatnagar, R.J., Arora, N., Sachidanand, S., Shahabuddin, M., Keister, D., Chauhan, V.S., 2003. Synthetic propeptide inhibits mosquito midgut chitinase and blocks sporogonic development of malaria parasite. Biochem. Biophys. Res. Commun. 304, 783–787.

Boguski, M., Lowe, T., Tolstoshev, C., 1993. dbEST—database for expressed sequence tags. Nat. Genet. 4, 332–333.

Boucher, Y., Doolittle, W.F., 2000. The role of lateral gene transfer in the esvolution of isoprenoid biosynthesis pathways. Mol. Microbiol. 37, 703–716.

Campbell, A., Mrazek, J., Karlin, S., 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc. Natl Acad. Sci. 96, 9184–9189.

Chen, W.-J., Fan-Chiang, M.-H., 2001. Directed migration of *Ascogregarina taiwanensis* (Apicomplexa:Lecudinidae) in its natural host *Aedes albopictus* (Diptera: Culicidae). J. Euk. Microbiol. 48, 537–541.

Claros, M., Vincens, P., 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur. J. Biochem. 241, 779–786.

Dacks, J.B., Marinets, A., Doolittle, W.F., Cavalier-Smith, T., Logsdon, J.M.J., 2002. Analyses of RNA polymerase II genes from free-living protists, Phylogeny, long branch attraction, and the eukaryotic big bang. Mol. Biol. Evol. 19, 830–840.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., Barrell, B., 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419, 498–511.

Hanaoka, H., Noda, T., Shirano, Y., Kato, T., Shibata, D., Tabata, S., Ohsumi, Y., 2002. Leaf senescence and starvation-induced chlorosis are accelerated by the disruption of an Arabidopsis autophagy gene. Plant Physiol. 129, 1181–1193.

Heintzelman, M.B., 2004. Actin and myosin in *Gregarina polymorpha*. Cell Motil & Cytosk 58, 83–95.

Henrissat, B., Davies, G., 2000. Glycoside hydrolases and glycosyltransferases. Families, modules and implications for genomics. Plant Physiol. 124, 1515–1519.

Hirt, R.P., Logsdon, J.M., Healy, B., Dorey, M.W., Doolittle, W.F., Embly, T.M., 1999. Microsporidia are related to fungi: evidence from the largest subunit of RNA polymerase II and other proteins. PNAS 96, 580–585.

Huang, J., Mullapudi, N., Sicheritz-Ponten, T., Kissinger, J., 2004. A first glimpse into the pattern and scale of gene transfer in Apicomplexa. Int. J. Parasitol. 34, 265–274.

Karlin, S., Cardon, L.R., 1994. Computational DNA sequence analysis. Ann. Rev. Microbiol. 48, 619–654.

Keeling, P.J., Doolittle, W.F., 1996. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. Mol. Biol. Evol. 13, 1297–1305.

Langer, R.C., Vinetz, J.M., 2001. Plasmodium ookinete-secreted chitinase and parasite penetration of the mosquito peritrophic matrix. Trends Parasitol. 17, 269–272.

Li, L., Brunk, B., Kissinger, J., Pape, D., Tang, K., Cole, R., Martin, J., Wylie, T., Dante, M., Fogarty, S., Howe, D., Liberator, P., Diaz, C., Anderson, J., White, M., Jerome, M., Johnson, E., Radke, J., Stoeckert, C.J., Waterston, R., Clifton, S., Roos, D., Sibley, L., 2003. Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. Genome Res. 13, 443–454.

Mackenzie, C., Walker, M.H., 1979. Bacteria-like structures in the endoplasm of *Gregarina garnhami* (Eugregarinida Protozoa). Cell Tissue Res. 202, 33–39.

Maeda, I., Kohara, Y., Yamamoto, M., Sugimoto, A., 2001. Large-scale analysis of gene function in Caenorhabditis elegans by high-throughput RNAi. Curr. Biol. 11, 171–176.

Moriya, S., Tanaka, K., Ohkuma, M., Sugano, S., Kudo, T., 2001. Diversification of the microtubule system in the early stage of eukaryote evolution: elongation factor 1 alpha and alpha-tubulin protein phylogeny of termite symbiotic oxymonad and hypermastigote protists. J. Mol. Evol. 52, 6–16.

Schrevel, J., Philippe, M., 1993. The gregarines. Parasitic Protozoa 4, 133–245.

Stiller, J.W., Duffield, E.C.S., Hall, B.D., 1998. Amitochondriate amoebae and the evolution of DNA-dependent RNA polymerase II. Proc. Natl Acad. Sci. 95, 11769–11774.

Wang, C., Klionsky, D., 2003. The molecular mechanism of autophagy. Mol. Med. 9, 65–76.