Published in final edited form as:

Int J Parasitol. 2014 October 1; 44(11): 837–845. doi:10.1016/j.ijpara.2014.06.008.

Regulation of gene expression in the protozoan parasite *Entamoeba invadens* identification of core promoter elements and promoters with stage-specific expression patterns

Dipak Manna^a, Gretchen M. Ehrenkaufer^a, and Upinder Singh^{a,b,*}

^aDivision of Infectious Diseases, Department of Internal Medicine, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, California, USA 94305

^bDepartment of Microbiology and Immunology, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, California, USA 94305

Abstract

Developmental switching between life-cycle stages is a common feature among many pathogenic organisms. *Entamoeba histolytica* is an important human pathogen and is a leading parasitic cause of death globally. During its life cycle, *Entamoeba* converts between cysts (essential for disease transmission) and trophozoites (responsible for tissue invasion). Despite being central to its biology, the triggers that are involved in the developmental pathways of this parasite are not well understood. In order to define the transcriptional network associated with stage conversion we used *Entamoeba invadens* which serves as a model system for *Entamoeba* developmental biology, and performed RNA sequencing at different developmental time points. In this study RNA-Seq data was utilized to define basal transcriptional control elements as well as to identify promoters which regulate stage-specific gene expression patterns. We discovered that the 5' and 3' untranslated regions of *E. invadens* genes are short, a median of 20 nucleotides (nt) and 26 nt respectively. Bioinformatics analysis of DNA sequences proximate to the start and stop codons identified two conserved motifs: (i) *E. invadens* Core Promoter Motif - GAAC-Like (*Ei*CPM-GL) (GAACATAAAA), and (ii) *E. invadens* 3' - U-Rich Motif (*Ei*3'-URM) (TTTGTT) in the 5' and 3' flanking regions, respectively. Electrophoretic mobility shift assays demonstrated that both motifs specifically bind nuclear protein(s) from *E. invadens* trophozoites. Additionally, we identified select genes with stage-specific expression patterns and analyzed the ability of each gene promoter to drive a luciferase reporter gene during the developmental cycle. This approach confirmed three trophozoite-specific, four encystation-specific and two excystation-specific promoters. This work lays the framework for use of stage-specific promoters to express proteins of interest in a particular life-cycle stage, adding to the molecular toolbox for genetic manipulation of *E. invadens* and allowing further dissection of factors controlling *Entamoeba* developmental biology.

© 2014 Elsevier Ltd. on behalf of Australian Society for Parasitology Inc. All rights reserved.

*Corresponding author. S-143 Grant Building, 300 Pasteur Drive, Stanford, CA, USA 94305. Tel.: 1-650-723-4045; fax: 1-650-724-3892. usingh@stanford.edu (Upinder Singh).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

Developmental regulation; RNA-Seq; Untranslated regions (UTR); Core promoter motif; Stage-specific promoters

1. Introduction

The protozoan parasite *Entamoeba histolytica* is a unicellular human pathogen and one of the leading parasitic causes of death worldwide (Stanley, 2003; Salles et al., 2007). Infection with *E. histolytica* predominantly manifests as dysentery due to colonic invasion but can also present with hepatic abscesses due to liver invasion. The life cycle of *Entamoeba* consists of two stages, an environmentally resistant, dormant cyst and an invasive trophozoite form. Infection by *E. histolytica* begins with ingestion of cysts through contaminated food or drinking water. Excystation occurs in the small bowel where the trophozoite, the motile and disease-causing form, emerges (Haque et al., 2003).

In amebic pathogenesis, stage inter-conversion between trophozoites and cysts is essential for pathogen transmission and disease manifestation (Haque et al., 2003). However, the triggers regulating the developmental pathway are not well understood, largely due to the lack of a system for studying stage conversion in *E. histolytica* in vitro (Eichinger, 1997; Singh and Ehrenkaufer, 2009). Thus the reptilian parasite, *Entamoeba invadens*, for which highly efficient in vitro encystation and excystation protocols exist, has been utilized as a model system to study *Entamoeba* development (Vazquezdelara-Cisneros and Arroyo-Begovich, 1984; Eichinger, 1997). Using this system, it has been shown that a number of processes including lipid signaling and meiosis are developmentally regulated (Ehrenkaufer et al., 2013). Additionally, as encystation progresses, the levels of metabolites involved in glycolysis decrease drastically (De Cadiz et al., 2013). It will be important to further characterize this dynamic change in metabolic regulatory networks if we are to gain a better understanding of the biology of stage conversion in *Entamoeba*.

Understanding the molecular mechanisms underlying coordinated regulation of gene expression is a crucial component of characterizing developmental control. *Entamoeba histolytica* has a number of transcriptional control features that are different from those of higher eukaryotes including an unusual RNA polymerase that is resistant to α -amanitin (Lioutas and Tannich, 1995) and very short untranslated regions (UTRs) (Bruchhaus et al., 1993; Purdy et al., 1996). Structurally, *E. histolytica* genes contain three conserved elements in their core promoters: a putative TATA element (GTATTATAA) at approximately 30 nucleotides (nt) upstream of the transcription initiation site, a GAAC element (AATGAAC) with variable location in the core promoter, and an Initiator (Inr) element (AAAAATTCA) overlying the transcription initiation site (Purdy et al., 1996; Singh et al., 1997, 2002). Importantly, the GAAC element was found to control the rate and site of transcription initiation independent of the TATA element (Singh et al., 2002).

In order to define the transcriptional network associated with stage conversion, we recently performed RNA sequencing of *E. invadens* during the entire developmental cycle (both encystation and excystation) (Ehrenkaufer et al., 2013). The RNA-Seq methodology was

used to improve the accuracy of genome annotation and identify untranslated regions as was demonstrated in *E. histolytica* (Hon et al., 2013), *Tetrahymena* (Xiong et al., 2012) and *Plasmodium* (Otto et al., 2010). Additionally, transcriptome data can be analyzed by bioinformatic approaches to identify conserved regulatory motifs as previously reported in *Giardia lamblia* (Tolba et al., 2013), *Trypanosoma brucei* (Mao et al., 2009), *Tetrahymena thermophila* (Xiong et al., 2012) and *E. histolytica* (Zamorano et al., 2008).

We utilized RNA-Seq data from *E. invadens* developmental time points to characterize basal transcriptional control elements as well as to identify promoters which regulate stage-specific gene expression patterns. Our analysis revealed core promoter features that are conserved in *E. invadens* compared with *E. histolytica*: relatively short 5'-UTRs and 3'-UTRs and a core promoter motif GAAC-like (*Ei*CPM-GL) (GAACTACAAA) highly reminiscent of the GAAC element identified previously in *E. histolytica* (Singh et al., 1997, 2002; Singh and Rogers, 1998). In addition, we identified promoters that regulate reporter gene expression in a stage-specific manner in *E. invadens*. These trophozoite-specific, cyst-specific and excystation-specific promoters will contribute significantly to the molecular toolbox for *E. invadens* and allow functional dissection of the molecular pathways regulating *Entamoeba* development.

2. Materials and methods

2.1. *Entamoeba invadens* culture and induction of stage conversion

Entamoeba invadens strain IP-1 was axenically maintained in LYI-S-2 medium at 25°C (Clark and Diamond, 2002). To induce encystation, trophozoites were incubated in 47% LYI-LG medium, as described previously (Sanchez et al., 1994). To determine the encystation efficiency, trophozoites were incubated overnight in distilled water, and intact cells before and after water treatment were counted to calculate the percentage of resistant cysts. For excystation experiments, trophozoites were encysted for 72 h using the above method and any remaining trophozoites were lysed by incubating overnight (~16 h) in distilled water at 4°C (a step that both lyses trophozoites and enhances excystation). The remaining cysts were induced to excyst by incubating in LYI-LG with 1 mg/ml of bile, 40 mM sodium bicarbonate, 1% glucose and 10% serum for 2 h or 8 h as previously described (Mitra et al., 2010).

2.2. Plasmid construction

We analyzed *E. invadens* RNA-Seq data (Ehrenkaufer et al., 2013) and selected transcripts that showed stage-specific expression patterns during development. Categories of stage-specific expression patterns are: trophozoite-specific (three genes, T1, T2 and T3, defined as expressed in trophozoites but with low or no expression in all cyst and excystation samples); cyst-specific (four genes, C1, C2, C3 and C4, defined as expressed in cysts at 24 h encystation but with low or no expression in trophozoites); and excystation-specific (two genes, Ex1 and Ex2, defined as expressed in excystation but with no or low expression in cysts) (Fig. 1). To make the stage-specific promoter constructs we used a p*Ei*-CK-myc-Luc vector, which was previously used to develop transfection methods in *E. invadens* (Ehrenkaufer and Singh, 2012). We replaced the 5'-casein kinase II (CK) promoter region

with individual stage-specific promoter regions. For each gene, approximately 500 nt upstream from the start codon were designated as the promoter region; in each instance, we incorporated the first six amino acids of the endogenous gene (which has previously been demonstrated to be important for optimal gene expression in *Entamoeba*) to an in-frame fusion with luciferase (Purdy et al., 1996; Ehrenkaufer and Singh, 2012). The 3' -regulatory regions in all constructs were from the CK gene, and a construct without any 5' -regulatory region served as a negative control. Three trophozoite-specific (EIN_057870, EIN_060140 and EIN_060150, named T1, T2 and T3, respectively), four cyst-specific (EIN_166430, EIN_147320, EIN_053400 and EIN_099680, named C1, C2, C3 and C4, respectively) and two excystation-specific (EIN_212890 and EIN_239520, named Ex1 and Ex2, respectively) promoter constructs were generated for this study. The primers used for this study are listed in Table 1.

2.3. Generation of transgenic parasites and luciferase assays

Entamoeba invadens strain IP-1 trophozoites were transfected by electroporation as previously described (Ehrenkaufer and Singh, 2012). Briefly, mid-log phase trophozoites were chilled, harvested, washed once with 1X PBS and resuspended in 0.4 ml of ZM buffer (132 mM NaCl, 8 mM KCl, 8 mM Na₂HPO₄, 1.5 mM KH₂PO₄, 0.5 mM magnesium acetate, 0.09 mM CaCl₂). Cells were electroporated using 50 µg of DNA, transferred into LYI-S-2 media in 15 ml glass tubes and left to recover at 25°C. Following a standard selection protocol, the G418 concentration was increased stepwise until stably transfected parasite lines at 40 µg/ml were obtained.

Luciferase assays were performed a minimum of three times as described earlier (Morf et al., 2013). Briefly, trophozoites were chilled, harvested and washed once with 1X PBS, re-suspended in 1X lysis buffer (Luciferase Assay System, #E1500, Promega, USA) complemented with protease inhibitors (1x HALT protein inhibitor cocktail, 1x E64), incubated on ice for 30 min to lyse trophozoites and spun at 20,800 g for 15 min. Cysts or excysted cells were lysed by sonication (five pulses at 15 amp for 15 s). Protein concentration was measured by a Bradford assay. A total of 30 µg of protein was added to luciferase reagent (Promega) and luciferase activity was measured by a Luminometer Monolight 2010.

2.4. Bioinformatic analysis of RNA-Seq data to identify UTRs and polyA addition sites

In order to determine the 5' and 3' UTR lengths we analysed the recent RNA-Seq data (Ehrenkaufer et al., 2013) using Bedtools (Quinlan and Hall, 2010). Bam files containing aligned sequences from 14 libraries, covering seven developmental timepoints, were merged and converted to Bed format. First, RNA-Seq reads were intersected with mRNA using intersectBed to select only those RNA-Seq reads that overlap with an annotated mRNA. Next, subtractBed was used to subtract the mRNA sequence from RNA-Seq data and closestBed tools was used to specify the shortest distance from the start or end of an mRNA to obtain the 5' and 3' UTR. The analysis was done for all annotated genes. Overall, a total of 6,287 transcripts had a significant number of reads mapped to either the 5' or 3' UTR. Because the RNA-Seq data gives poorer coverage at the 5'-end of a gene and since this could affect determination of the 5' UTR size from long genes we further analyzed the UTR

data for genes that are 500 bp in size as well as for those that are 1,000 bp in size. We also analyzed the 5' UTR length in select genes with high mRNA expression (Fragments Per Kilobase of exon per Million mapped reads (FPKM) values > 25,000) since those will have better 5' sequence coverage. Determination of the polyA addition site was done using a method adapted from Hon et al. (2013). RNA-seq data from 14 libraries representing seven developmental timepoints (Ehrenkauf et al., 2013) were merged and mapped to the *E. invadens* genome using bowtie (Langmead et al., 2009) with the parameters: -v 3 --trim3 14. All unmapped reads were converted from colorspace to FASTA and reads containing either 4 - 5 As at the 3' end or 4 - 5 Ts at the 5' end (which may be generated from the region spanning the polyA start site) were extracted and trimmed to remove the ends using a custom python script. These reads were remapped to the genome and the resultant mapping file filtered to remove reads that overlapped completely with annotated genes or that contained runs of >20 As or Ts. Loci with multiple reads mapping to them were identified and the distance to the 3'-end of the nearest gene was calculated. Clusters that were >200 bp from the closest 3'-end were considered artifacts and removed from the final analysis.

In order to identify potential regulatory motifs in the core promoter or immediate downstream regions of a gene, the MEME program (<http://meme.nbcr.net/meme/>) was used as previously described (Hackney et al., 2007; Pearson et al., 2013). In brief, 150 nt upstream from the ATG start codon of all annotated transcripts were analyzed to identify conserved core promoter motifs. To identify conserved motifs in the 3'-regulatory regions, 150 nt downstream from the stop codon were analyzed. All *E. invadens* transcripts were randomly divided into three subsets with ~ 4,000 genes in each subset. MEME was performed with the parameters: -mod zoops -minw 4 -maxw 10 -minsites 50 -nmotfs 30. Given that the genes were randomly divided into three subsets, motifs that predominate in all subsets are likely to have biological significance. Sequence logos were generated using the R package SeqLogo.

2.5. Electrophoretic mobility shift assays (EMSAs)

EMSAs were performed as previously described (Pearson et al., 2013). The oligonucleotides used in EMSAs were as follows (mutations underlined) – (i) *Ei*CPM-GL wild type (WT) sense: 5' -acgtcagtAGGCGAACTACAAAAGATcagt- 3', (ii) *Ei*CPM-GL WT anti-sense: 5' -acgtactgATCTTTTGTAGTTCGCCTactg- 3'; (iii) *Ei*CPM-GL Mutant sense: 5' -acgtcagtAGGCCGATAATAGTAGATcagt- 3', (iv) *Ei*CPM-GL Mutant anti-sense: 5' -acgtactgATCTACTATTATCGGCCTactg- 3'; (v) *E. invadens* 3'- U-Rich Motif (*Ei*3'-URM) WT sense: 5' -acgtcagtTGCATTTGTTTAAAcagt- 3', (vi) *Ei*3'-URM WT anti-sense: 5' -acgtactgTTTAAACAAATGCAactg- 3'; (vii) *Ei*3'-URM Mutant sense: 5' -acgtcagtTGCACAGATATAAAcagt- 3', (viii) *Ei*3'-URM Mutant anti-sense: 5' -acgtactgTTTATATCTGTGCAactg- 3'. Complementary overlapping probes were annealed and labeled using [³²P] α-ATP and Klenow fragment (Invitrogen, USA) (Hackney et al., 2007). Binding reactions were carried out as described previously (Pearson et al., 2013). Briefly, binding reaction was set in a total volume of 20 µl, which included 2 µl of 10X EMSA binding buffer (10 mM Tris-HCl, pH 7.9, 50 mM NaCl, 1 mM EDTA, 3% glycerol, 0.05% milk powder and 0.05 mg of bromophenol blue), 5 µg of nuclear extract, 2 µg of poly (dI-dC), and 50 fmol of labeled probe. The binding reaction mix was incubated for 30 min at

room temperature, and samples were loaded onto a 9% non-denaturing polyacrylamide gel and run for 3 h. The gel was fixed, dried and exposed to a phosphor screen. Gels were imaged using a Personal Molecular Imager (PMI) System with Quantity One software, (Bio-Rad, USA).

3. Results

3.1. Determination of the UTR length and polyA addition sites

Understanding the core promoter architecture is a first step in dissecting mechanisms that regulate gene expression. Using the genome-wide transcriptomic RNA-Seq data in *E. invadens*, we determined the UTR lengths and polyA addition sites. Our RNA-Seq data from 14 libraries, covering seven developmental timepoints, were merged. We observed that the RNA-Seq reads map mostly to annotated mRNA sequences; however, some regions of the genome have significant numbers of mapped reads despite having no annotated genes. To simplify our analysis, all regions with mapped RNA-Seq reads that did not overlap with an annotated mRNA were removed from the dataset. Next, in order to determine the length of the 5' and 3' UTRs, we subtracted the annotated coding region of all genes from the intersected RNA-Seq data set using BEDTools. Fig. 2A is an example of the 5' and 3' UTR lengths of a transcript (EIN_160240), which codes for a hypothetical protein. The RNA-Seq coverage is longer than the annotated coding sequence and after subtracting coding sequence from the RNA-Seq data, we determined the 5' and 3' UTR length of EIN_160240 transcript as nine nt and 21 nt, respectively (Fig. 2A).

We observed that the cumulative FPKM values of individual transcripts in the data set needed to be relatively high (> 20 FPKM) in transcripts, in order for our approach to reliably detect a UTR. Our RNA-Seq data identified 2,746 transcripts with a detectable 5' UTR (i.e. 24% of total transcripts) with a median of 20 nt and mode of eight nt (Fig. 2B). Analysis of the 3' UTR revealed 3,618 transcripts with a detectable 3' UTR (i.e. 31% of total transcripts). The lengths of the 3' UTR were also relatively short (mostly within < 50 nt) with a median of 26 nt and mode of 19 nt (Fig. 2C). An explanation for the overall fewer genes with detectable 5' UTR is that the 5' region may be a great distance from the 3' poly-A sites where cDNA generation is initiated (Fig. 2B, C). In order to see whether variability in transcript length affects the observed length of 5' UTR, we used genes with variable lengths (<1 Kb and <500 bp), which should have better RNASeq coverage at their 5' regions. The median length of 5' UTRs was observed as 23 nt and 29 nt in transcripts which are <1 Kb and <500 bp, respectively (Supplementary Fig. S1). Overall, the UTR length distribution revealed that 84% of the 5' UTR and 90% of the 3' UTR were shorter than 100 nt. We also analyzed genes with high expression levels and found that the UTR length in the highly expressed gene subset was similar to what was determined above (data not shown). Thus, the *E. invadens* UTRs are relatively short in length compared with the metazoan average of 60 - 80 nt (Kozak, 1984).

These data of short 5' and 3' UTRs confirm what has been noted in *E. histolytica*, which also has relatively short 5' UTRs (median 11 nt) and 3' UTRs (median 21 nt) (Purdy et al., 1996; Hon et al., 2013). However, longer 5' and 3' UTRs have been noted for specific amebic genes as described for the 280 nt 5' UTR of the *E. histolytica* stress-sensitive protein

(Satish et al., 2003) and the 264 nt 3' UTR of the *EhP*-glycoprotein 5 gene (Lopez et al., 2000). Thus, exceptions to the observation of short 5' and 3' regulatory regions do occur, an observation we also made in *E. invadens* (data not shown).

Mapping of the polyA addition sites of *E. invadens* transcripts confirmed the observed short UTR length. For this analysis, reads spanning the 3' end of the transcript and the start of the polyA tail were identified and the proximity to the translation stop site was calculated. Due to the limited number of reads that met our criteria (i.e. to have sufficient sequence reads both upstream and downstream of the polyA site to be conclusively identified and remapped) only 657 transcripts (which had multiple reads mapped to their putative polyA site) were used for this analysis. We found that the majority of polyA start sites were found within 20 – 30 nt of the end of the coding region, with the median distance being 24 nt and the most common distance being 21 nt; Fig. 2D shows the distribution of calculated distances. This confirms the short 3' UTR length we demonstrated above for *E. invadens* and is similar to what was found in *E. histolytica*, where the median position of polyA sites was found to be 21 nt from the transcription end site (Hon et al., 2013).

3.2. Identification of core promoter regulatory regions

To determine whether specific sequences were conserved in the 5' and 3' core promoter regions we utilized MEME to analyze all annotated *E. invadens* transcripts to identify conserved motifs. All annotated transcripts were randomly subdivided into three subsets, each with ~4,000 transcripts; identifying a motif which predominates in all three subsets strengthens the likelihood of its biological significance. To identify conserved core promoter elements, we analyzed 150 nt upstream from the ATG start codon for all annotated transcripts. Earlier studies identified that the *E. histolytica* core promoter contains three conserved elements: (i) a TATA element (GTATTTAAA), (ii) a GAAC element (AATGAACT) with variable location between the TATA and Inr, and (iii) an Inr (AAAAATTCA), at the site of transcription initiation (Purdy et al., 1996; Singh et al., 1997, 2002). Each of these elements was important in gene regulation, as demonstrated by mutational analysis, with the GAAC element having the greatest effect on gene expression (Purdy et al., 1996; Singh et al., 1997). Additionally, it was demonstrated that both the TATA and GAAC elements could direct transcription start sites independently of each other (Singh et al., 2002).

Our analysis of all *E. invadens* promoter regions identified 13 motifs which showed more than 50 occurrences in each dataset (Supplementary Table S1). Among these motifs one motif predominated, with a consensus sequence of GAAC TACAAA (Fig. 3A). We named this *Ei*CPM-GL. This motif was present in the 5' flanking regions of ~31% of the total 11,549 transcripts (total of 3,635 occurrences), resided largely within 30 nt upstream of the start codon (Fig. 3B), and appears highly reminiscent of the *E. histolytica* GAAC sequence (AATGAACT). The *E. invadens Ei*CPM-GL motif could represent a combination of two motifs which have previously been noted in *E. histolytica* core promoters: GAAC (AATGAACT) and the Inr element (AAAAATTCA), especially with apparent overlap with the highly conserved AAA region of the Inr element (Purdy et al., 1996; Singh et al., 1997, 2002). In *E. histolytica*, the GAAC core promoter element regulates transcriptional

activation and protein complex assembly, and directs the transcription start site (Purdy et al., 1996; Singh et al., 1997, 2002). Thus, it is possible that the *Ei*CPM-GL motif may also mediate transcriptional control in *E. invadens*.

Our analysis of *E. invadens* core promoters did not identify a motif similar to TATA or Inr elements noted in *E. histolytica* (Supplementary Table S1). In order to exclude trivial technical reasons, we performed a similar core promoter analysis of *E. histolytica* genes. Bioinformatics analysis of 4,000 randomly selected *E. histolytica* promoters identified an identical *Ei*CPM-GL motif as identified in *E. invadens*. The motif was present in 2,241 genes (56%) out of the 4,000 promoters analyzed, a similar ratio as was seen in *E. invadens*. However, our analysis only identified 56 genes (1% of total) with TATA-like (ATATTTAAAG) sequences in *E. histolytica* promoter regions. Additionally, the bioinformatic analysis did not identify a conserved Inr element (data not shown). It is important to note that the TATA, GAAC and Inr elements were initially identified in *E. histolytica* by aligning only 16 amebic genes (including *hgl5*, *hgl2*, actin and *Ferdx*) at their site of transcription initiation (Purdy et al., 1996). One explanation for the lack of identification of a TATA element by the large-scale bioinformatic analysis is that it may be hard to find an AT rich motif in a highly AT rich genome using MEME analysis. Another possibility is that the TATA element is more highly conserved in promoters of highly expressed genes (*hgl5*, *hgl2*, actin and *Ferdx* genes initially used to identify the promoter sequences were a select set of genes with high expression levels). A TATA box binding protein (*Eh*TBP) has been identified in *E. histolytica* and genes which contain different TATA elements bind recombinant *Eh*TBP in vitro at different affinities (de Dios-Bravo et al., 2005). This suggests that the *Entamoeba* TATA box is not well conserved and shows more relaxed binding specificity with *Eh*TBP than described in other systems (Wobbe and Struhl, 1990; Patikoglou et al., 1999). A TBP was identified in the *E. invadens* genome (Ehrenkaufer et al., 2013), and thus a version of a TATA box likely exists in *E. invadens*. The previously identified *E. histolytica* Inr element had the weakest sequence conservation and the least effect on gene expression, and was largely identified by mapping the transcription start site and looking for conserved motifs in this region (Purdy et al., 1996). Thus, it may be more difficult to identify in *E. invadens* using large-scale MEME analyses. Interestingly, the currently identified *Ei*CPM-GL motif in both in *E. invadens* and *E. histolytica* promoters appears to be a combination of previously reported GAAC and Inr elements.

To identify 3' regulatory elements, we analyzed 150 nt downstream from the stop codon of 11,549 genes. Five motifs, which showed more than 50 occurrences in each dataset, were identified (Supplementary Table S2). The motif that predominated in all subsets was TTTGTT, the *Ei*3'-URM, which largely resided within 50 nt downstream of the stop codon (Fig. 3C, D). This motif occurred in 5,152 genes (~44.6% of the total 11,549 transcripts) and was thus highly abundant. In *E. histolytica*, a similar *Ei*3'-URM motif was previously noted as a U-rich motif which resides downstream of poly-A sites (Zamorano et al., 2008).

3.3. Conserved EiCPM-GL and Ei3'-URM motifs specifically bind *E. invadens* nuclear protein(s)

The motif *Ei*CPM-GL (GAACTACAAA), which is conserved in the upstream region of almost 31% of *E. invadens* transcripts may have biological significance in gene regulation. In *E. histolytica* the GAAC core promoter element regulates gene expression and binds *E. histolytica* nuclear protein(s) (Singh et al., 2002). In order to determine whether the CPM we identified in *E. invadens* binds nuclear proteins we performed EMSA. A radiolabeled *Ei*CPM-GL motif was incubated with nuclear extract prepared from *E. invadens* trophozoites. EMSA analysis demonstrated that the consensus probe binds to protein(s) present in nuclear extract and that this interaction decreased in the presence of two- or 10-fold excess of cold *Ei*CPM-GL WT probe, indicating the specificity of the interaction (Fig. 4A). Mutations of the *Ei*CPM-GL motif were created by changing seven nucleotides of the consensus sequence and tested for DNA-protein binding. Compared with the WT competitor, this mutant motif, *Ei*CPM-GL-Mutant (GAACTACAAA changed to CGATAATAGT), was unable to compete in an EMSA assay, suggesting that these core nucleotides are important for protein-DNA binding. The *Ei*CPM-GL motif has similar binding affinities, as shown by competition experiments, and similar residues critical for binding, as shown by mutational analysis, as the *E. histolytica* GAAC element (Singh et al., 2002). It was previously demonstrated that the GAAC region is necessary for higher-order nuclear protein complex assembly in *E. histolytica* and similar roles may also occur in *E. invadens* (Singh et al., 2002).

To determine whether the *Ei*3'-URM motif (TTTGTT) binds *E. invadens* nuclear protein we performed EMSA analysis. Our results showed that the WT *Ei*3'-URM motif binds to protein(s) present in crude nuclear extract and that this interaction decreased in the presence of a 10 or 100-fold excess of cold WT *Ei*3'-URM probe, indicating specificity of the interaction (Fig. 4B). Mutation of the WT *Ei*3'-URM motif (TTTGTT to CAGATA) was unable to compete by EMSA binding, thus confirming the specificity of the interaction. It was previously reported that the *E. histolytica* cleavage stimulation factor (*Eh*CstF) protein complex could bind U/GU-rich motif and facilitate pre-mRNA 3' end processing (Lopez-Camarillo et al., 2005). *Entamoeba invadens* has the CstF homolog (53% identical to *Eh*CstF) and may be involved in similar functions. The *Ei*3'-URM motif is present in 44.6% of the total 11,549 transcripts and is quite similar to the U/GU-rich motif described previously, and thus may have important roles in mRNA 3' end processing and gene regulation.

3.4. Identification of stage-specific promoters and validation by luciferase reporter assay

The transcriptomic profile of *E. invadens* during the developmental stages of encystation and excystation was recently reported (De Cadiz et al., 2013; Ehrenkaufer et al., 2013). In order to determine whether promoters have stage-specific expression patterns, we utilized the RNASeq data to identify stage-specific genes and analyze their individual promoter regions. Such stage-specific promoter constructs could be useful to achieve regulated expression of a protein of interest. We identified transcripts with stage-specific expression patterns: highly expressed only in trophozoites (T1, T2, T3), during encystation (C1, C2, C3, C4), or during excystation (Ex1, Ex2) (Fig. 1). We confirmed that the expression patterns of

the trophozoite and cyst-specific genes also matched the *E. invadens* microarray data (De Cadiz et al., 2013). To make stage-specific promoter constructs we considered several parameters. First, we chose stage-specific transcripts that had consistent expression patterns in both biological RNA-Seq replicates. Second, we did not use a promoter if there was an adjacent gene within 500 bp. For trophozoite-specific genes, we picked genes that were expressed in trophozoites but with low or no expression in all cyst and excystation samples. To generate cyst-specific constructs we selected transcripts that showed increased expression at 24 h of encystation but remained low in trophozoites and at early stages of encystation (8 h), as this timepoint may have noisy data from a generalized stress response. For excystation-specific genes, we chose genes that expressed in excystation but with no or low expression in cysts. Promoter regions (~500 nt) of differentially expressed transcripts were cloned upstream of the luciferase reporter gene into the pEi-myc-Luc vector. The 3' regulatory region in all constructs was the 3' UTR of the CKII gene (EIN_083650) (Ehrenkaufer and Singh, 2012). Promoter constructs were transfected into *E. invadens* trophozoites and stable transfectants maintained at 40 µg/ml of G418.

Luciferase assays show all three trophozoite-specific promoters (T1, T2 and T3) are highly active in trophozoites but have low expression in cysts (Fig. 5B), validating the RNA-Seq data (Fig. 5A). Among the three trophozoite-specific constructs, T2 had the highest FPKM values and had high luciferase activity. However, the FPKM values of T1 and T3 do not completely reflect in the level of luciferase activity. On the contrary, other stage-specific promoters (cyst-specific or excystation-specific promoters) had low activity in trophozoites. All four cyst-specific promoter constructs (C1, C2, C3 and C4) showed higher activity in 18 h cysts compared with trophozoites (Fig. 5D). C2 and C4 have statistically higher expression in cysts compared with trophozoites, whereas C1 and C3 are higher in cysts although not statistically significant. The excystation-specific promoters (Ex1, Ex2) showed low activity in trophozoites and high activity at 8 h of excystation (Fig. 5F). Overall, the promoters we analyzed regulated luciferase expression in a manner that matched the RNA-Seq data and indicates that at least a subset of stage-specific gene expression is controlled by promoter activity.

4. Discussion

Regulation of gene expression and stage conversion is poorly understood in *Entamoeba*. Understanding the mechanisms underlying coordinated regulation of gene expression during developmental stages is crucial as they contribute to the parasite's pathogenicity. This study utilizes the global analysis of RNA-Seq data from seven developmental timepoints of *E. invadens*, including encystation and excystation (Ehrenkaufer et al., 2013), to provide an overview of transcriptional regulatory elements in *E. invadens*. Our work identified some common basal transcriptional features (short UTRs and CPMs) between *E. histolytica* and *E. invadens*, as well as identifying stage-specific promoters which can be used for regulated protein expression and functional characterization of the molecular triggers that control parasite developmental biology.

In *E. histolytica* the 5' and 3' UTRs are short (Bruchhaus et al., 1993), with the 5' UTR having an average length of 11 nt compared with a metazoan average of 60 – 80 nt (Kozak,

1984). The 3' UTR is relatively longer with an average size of 21 – 33 nt (Bruchhaus et al., 1993; Hon et al., 2013). We now demonstrate from RNA-Seq data of *E. invadens* that 5' and 3' UTRs are relatively short, similar to other enteric protozoan parasites, with a median length of 20 nt and 26 nt, respectively. Another protozoan parasite, *G. lamblia*, also has short 5' and 3' UTRs ranging from 0 – 14 nt and 10 – 30 nt, respectively (Adam, 2000). The UTR regions of *Trichomonas vaginalis* range from 5' UTRs of 6 – 12 nt and 3' UTRs of 10 – 30 nt (Quon et al., 1994; Fuentes et al., 2012). Given the shared features of short UTRs in these enteric parasites, they may share common molecular mechanisms of translation initiation and RNA stability.

Promoter organization of *E. histolytica* is unusual compared with other protozoan promoters, with three core promoter elements: a non-consensus TATA element, a GAAC element and an unusual Inr element (Purdy et al., 1996; Singh et al., 1997, 2002; Singh and Rogers, 1998). Our genome-wide bioinformatic analysis in *E. invadens* identifies a single *Ei*CPM-GL motif (GAAC TACAAA), which is likely a combination of the previously reported GAAC and Inr regions. This motif binds *E. invadens* nuclear protein(s) and is likely important in basal transcriptional control given its highly abundant nature (present in 31% of promoters). However, we could not find the amebic consensus TATA-box or Inr element in *E. invadens*. Given that a TBP has been cloned in *E. histolytica* and the gene coding for TBP is present in *E. invadens*, our inability to identify a TATA box by bioinformatic analyses means that a technical rather than biological explanation is most likely. One explanation is that it might be difficult to find an AT-rich TATA-like motif in an AT-rich genome using MEME analysis. Another possibility is that the *Ei*CPM-GL motif binds a multi-protein complex that contains TBP, but that the DNA binding of the complex is mediated by the *Ei*CPM-GL motif and its binding protein(s). Identification of the protein(s) that bind to the GAAC motifs in *E. histolytica* and *E. invadens* will help to resolve this issue.

The 3' UTR can mediate mRNA turnover and the control of gene transcript expression. In *E. histolytica*, at the 3' UTR a conserved pentanucleotide (TA^A/T^TTT) motif (Bruchhaus et al., 1993) and an AAUUA motif upstream of poly (A) sites have been reported (Hon et al., 2013). Analysis of 3' UTR regions of *E. invadens* transcripts identified a consensus *Ei*3'-URM motif (TTTGTT) that is present in almost 44.6% of the total transcripts. This motif was identified as a URM in *E. histolytica* transcripts but its function is not yet characterized (Zamorano et al., 2008). Our EMSA analysis showed that this *Ei*3'-URM motif specifically binds to *E. invadens* nuclear protein(s), indicating that it is likely to be biologically significant in gene regulation.

To explore stage-specific expression via specific promoter sequences, luciferase reporter constructs were analyzed. In *G. lamblia*, an encystation-specific promoter has been identified, namely the promoter region of cyst wall protein (cwp2), that has high expression during encystation and is regulated by multiple cis acting elements (Davis-Hayman et al., 2003). We took a similar strategy by selecting transcripts, which showed stage-specific expression, from recent RNA-Seq data (Ehrenkaufer et al., 2013). Given that multiple mechanisms for regulating gene expression have been described in *Entamoeba* (transcription factor, epigenetic control and small RNA regulation) (Mirelman et al., 2008; Singh and Ehrenkaufer, 2009; Zhang et al., 2011; Morf et al., 2013), it was not clear whether stage-

specific mRNA patterns would be recapitulated in stage-specific promoter activity. For the multiple promoters analyzed, we found concordant regulation of luciferase expression in a stage-specific manner, thus validating the RNA-Seq data. This implies that, at least for this subset of genes, expression is controlled by upstream cis-regulatory elements. The promoters that we tested represent a very small number from the entire genome and it is not clear whether our findings have broad generalizability: i.e. is the majority of stage-specific gene expression controlled by regulated promoter activity or by other gene regulation mechanisms? However, these stage-specific promoter constructs will allow expression of proteins of interest in a particular stage and will thus be useful for studying the developmental cascade of *Entamoeba*. Future studies to identify the motifs within the promoters that impart stage-specific promoter activity will be useful and can help to identify transcription factors that control expression of developmentally regulated genes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank all members of the Singh laboratory for valuable suggestions and technical assistance. This research was supported by the National Institutes of Health, USA grant AI-094887 to US.

References

- Adam RD. The *Giardia lamblia* genome. *Int. J. Parasitol.* 2000; 30:475–484. [PubMed: 10731570]
- Bruchhaus I, Leippe M, Lioutas C, Tannich E. Unusual gene organization in the protozoan parasite *Entamoeba histolytica*. *DNA Cell Biol.* 1993; 12:925–933. [PubMed: 8274224]
- Clark CG, Diamond LS. Methods for cultivation of luminal parasitic protists of clinical importance. *Clin. Microbiol. Rev.* 2002; 15:329–341. [PubMed: 12097242]
- Davis-Hayman SR, Hayman JR, Nash TE. Encystation-specific regulation of the cyst wall protein 2 gene in *Giardia lamblia* by multiple cis-acting elements. *Int. J. Parasitol.* 2003; 33:1005–1012. [PubMed: 13129521]
- De Cadiz AE, Jeelani G, Nakada-Tsukui K, Caler E, Nozaki T. Transcriptome analysis of encystation in *Entamoeba invadens*. *PloS one.* 2013; 8:e74840. [PubMed: 24040350]
- de Dios-Bravo G, Luna-Arias JP, Riveron AM, Olivares-Trejo JJ, Lopez-Camarillo C, Orozco E. *Entamoeba histolytica* TATA-box binding protein binds to different TATA variants in vitro. *FEBS J.* 2005; 272:1354–1366. [PubMed: 15752353]
- Ehrenkaufer GM, Singh U. Transient and stable transfection in the protozoan parasite *Entamoeba invadens*. *Mol. Biochem. Parasitol.* 2012; 184:59–62. [PubMed: 22561071]
- Ehrenkaufer GM, Weedall GD, Williams D, Lorenzi HA, Caler E, Hall N, Singh U. The genome and transcriptome of the enteric parasite *Entamoeba invadens*, a model for encystation. *Genome Biol.* 2013; 14:R77. [PubMed: 23889909]
- Eichinger D. Encystation of entamoeba parasites. *BioEssays.* 1997; 19:633–639. [PubMed: 9230696]
- Fuentes V, Barrera G, Sanchez J, Hernandez R, Lopez-Villasenor I. Functional analysis of sequence motifs involved in the polyadenylation of *Trichomonas vaginalis* mRNAs. *Eukary. Cell.* 2012; 11:725–734.
- Hackney JA, Ehrenkaufer GM, Singh U. Identification of putative transcriptional regulatory networks in *Entamoeba histolytica* using Bayesian inference. *Nucl. Acids Res.* 2007; 35:2141–2152. [PubMed: 17355990]
- Haque R, Huston CD, Hughes M, Hout E, Petri WA Jr. Amebiasis. *New Engl. J. Med.* 2003; 348:1565–1573. [PubMed: 12700377]

- Hon CC, Weber C, Sismeiro O, Proux C, Koutero M, Deloger M, Das S, Agrahari M, Dillies MA, Jagla B, Coppee JY, Bhattacharya A, Guillen N. Quantification of stochastic noise of splicing and polyadenylation in *Entamoeba histolytica*. *Nucl. Acids Res.* 2013; 41:1936–1952. [PubMed: 23258700]
- Kozak M. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucl. Acids Res.* 1984; 12:857–872. [PubMed: 6694911]
- Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009; 10:R25. [PubMed: 19261174]
- Lioutas C, Tannich E. Transcription of protein-coding genes in *Entamoeba histolytica* is insensitive to high concentrations of alpha-amanitin. *Mol. Biochem. Parasitol.* 1995; 73:259–261. [PubMed: 8577336]
- Lopez C, Marchat LA, Luna-Arias JP, Orozco E. An initial characterization of the 3' untranslated region of the EhPgp5 mRNA in *Entamoeba histolytica*. *Arch. Med. Res.* 2000; 31:S282–S284. [PubMed: 11070317]
- Lopez-Camarillo C, Orozco E, Marchat LA. *Entamoeba histolytica*: comparative genomics of the pre-mRNA 3' end processing machinery. *Exp. Parasitol.* 2005; 110:184–190. [PubMed: 15955310]
- Mao Y, Najafabadi HS, Salavati R. Genome-wide computational identification of functional RNA elements in *Trypanosoma brucei*. *BMC Genomics.* 2009; 10:355. [PubMed: 19653906]
- Mirelman D, Anbar M, Bracha R. Epigenetic transcriptional gene silencing in *Entamoeba histolytica*. *IUBMB Life.* 2008; 60:598–604. [PubMed: 18493998]
- Mitra BN, Pradel G, Frevert U, Eichinger D. Compounds of the upper gastrointestinal tract induce rapid and efficient excystation of *Entamoeba invadens*. *Int. J. Parasitol.* 2010; 40:751–760. [PubMed: 20018192]
- Morf L, Pearson RJ, Wang AS, Singh U. Robust gene silencing mediated by antisense small RNAs in the pathogenic protist *Entamoeba histolytica*. *Nucl. Acids Res.* 2013; 41:9424–9437. [PubMed: 23935116]
- Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, Bohme U, Lemieux J, Barrell B, Pain A, Berriman M, Newbold C, Llinas M. New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Mol. Microbiol.* 2010; 76:12–24. [PubMed: 20141604]
- Patikoglou GA, Kim JL, Sun L, Yang SH, Kodadek T, Burley SK. TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev.* 1999; 13:3217–3230. [PubMed: 10617571]
- Pearson RJ, Morf L, Singh U. Regulation of H₂O₂ stress-responsive genes through a novel transcription factor in the protozoan pathogen *Entamoeba histolytica*. *J. Biol. Chem.* 2013; 288:4462–4474. [PubMed: 23250742]
- Purdy JE, Pho LT, Mann BJ, Petri WA Jr. Upstream regulatory elements controlling expression of the *Entamoeba histolytica* lectin. *Mol. Biochem. Parasitol.* 1996; 78:91–103. [PubMed: 8813680]
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010; 26:841–842. [PubMed: 20110278]
- Quon DV, Delgadillo MG, Khachi A, Smale ST, Johnson PJ. Similarity between a ubiquitous promoter element in an ancient eukaryote and mammalian initiator elements. *Proc. Nat. Acad. Sci. USA.* 1994; 91:4579–4583. [PubMed: 8183951]
- Salles JM, Salles MJ, Moraes LA, Silva MC. Invasive amebiasis: an update on diagnosis and management. *Expert Review Anti-infective Therapy.* 2007; 5:893–901.
- Sanchez L, Enea V, Eichinger D. Identification of a developmentally regulated transcript expressed during encystation of *Entamoeba invadens*. *Mol. Biochem. Parasitol.* 1994; 67:125–135. [PubMed: 7838173]
- Satish S, Bakre AA, Bhattacharya S, Bhattacharya A. Stress-dependent expression of a polymorphic, charged antigen in the protozoan parasite *Entamoeba histolytica*. *Infect. Immun.* 2003; 71:4472–4486. [PubMed: 12874327]
- Singh U, Ehrenkaufer GM. Recent insights into *Entamoeba* development: identification of transcriptional networks associated with stage conversion. *Int. J. Parasitol.* 2009; 39:41–47. [PubMed: 18938171]

- Singh U, Gilchrist CA, Schaenman JM, Rogers JB, Hockensmith JW, Mann BJ, Petri WA. Context-dependent roles of the *Entamoeba histolytica* core promoter element GAAC in transcriptional activation and protein complex assembly. *Mol. Biochem Parasitol.* 2002; 120:107–116. [PubMed: 11849710]
- Singh U, Rogers JB. The novel core promoter element GAAC in the hgl5 gene of *Entamoeba histolytica* is able to direct a transcription start site independent of TATA or initiator regions. *J. Biol. Chem.* 1998; 273:21663–21668. [PubMed: 9705300]
- Singh U, Rogers JB, Mann BJ, Petri WA Jr. Transcription initiation is controlled by three core promoter elements in the hgl5 gene of the protozoan parasite *Entamoeba histolytica*. *Proc. Nat. Acad. Sci. USA.* 1997; 94:8812–8817. [PubMed: 9238060]
- Stanley SL Jr. Amoebiasis. *Lancet.* 2003; 361:1025–1034. [PubMed: 12660071]
- Tolba ME, Kobayashi S, Imada M, Suzuki Y, Sugano S. *Giardia lamblia* transcriptome analysis using TSS-Seq and RNA-Seq. *PloS one.* 2013; 8:e76184. [PubMed: 24116096]
- Vazquezdelara-Cisneros LG, Arroyo-Begovich A. Induction of encystation of *Entamoeba invadens* by removal of glucose from the culture medium. *J. Parasitol.* 1984; 70:629–633. [PubMed: 6512629]
- Wobbe CR, Struhl K. Yeast and human TATA-binding proteins have nearly identical DNA sequence requirements for transcription in vitro. *Mol. Cell. Biol.* 1990; 10:3859–3867. [PubMed: 2196437]
- Xiong J, Lu X, Zhou Z, Chang Y, Yuan D, Tian M, Zhou Z, Wang L, Fu C, Orias E, Miao W. Transcriptome analysis of the model protozoan, *Tetrahymena thermophila*, using Deep RNA sequencing. *PloS one.* 2012; 7:e30630. [PubMed: 22347391]
- Zamorano A, Lopez-Camarillo C, Orozco E, Weber C, Guillen N, Marchat LA. In silico analysis of EST and genomic sequences allowed the prediction of cis-regulatory elements for *Entamoeba histolytica* mRNA polyadenylation. *Comp. Biol. Chem.* 2008; 32:256–263.
- Zhang H, Pompey JM, Singh U. RNA interference in *Entamoeba histolytica*: implications for parasite biology and gene silencing. *Future Microbiol.* 2011; 6:103–117. [PubMed: 21162639]

Highlights

- *Entamoeba invadens* core transcriptional features were identified
- Stage-specific promoters were identified and validated
- Regulated expression of proteins will be feasible
- Regulated expression of proteins will be an important addition to the genetic tool-box for this organism

Fig. 1.

Identification of stage-specific promoters in *Entamoeba invadens*. Schematic representations of *E. invadens* RNA sequencing data which showed time points (h) of development on the X-axis and expression level (Fragments Per Kilobase of exon per Million mapped reads (FPKM)) on the Y-axis. On the basis of RNA-Seq data we made several stage-specific promoter constructs including trophozoite-specific (T1, T2 and T3); cyst-specific (C1, C2, C3 and C4); and excystation-specific (Ex1 and Ex2). Troph, trophozoite.

Fig. 2.

Determination of the 5' and 3' untranslated region (UTR) lengths of *Entamoeba invadens* genes. (A) Representative schematic of the UTR length determination by subtracting mRNA length from the RNA-Seq data. The identified lengths of the 5'-UTRs and 3'-UTRs of the transcript EIN_160240 are nine nucleotides (nt) and 21 nt, respectively. (B) Distribution of the 5'-UTR length of *E. invadens* transcripts. (C) Distribution of the 3'-UTR length of *E. invadens* transcripts. (D) Distribution of polyA site position relative to the stop codon.

Fig. 3. Identification of *Entamoeba invadens* Core Promoter Motif GAAC-Like (*Ei*CPM-GL) and 3'-U-Rich Motif (*Ei*3'-URM). (A, C) MEME analysis identified a conserved *Ei*CPM-GL (GAACTACAAA) in the 5'-promoter region and a conserved *Ei*3'-URM (TTTGTT) in the 3' region of *E. invadens* transcripts, respectively. (B, D) *Ei*CPM-GL and *Ei*3'-URM are predominantly located within 30 nucleotides (nt) of the start (ATG) and stop codon, respectively.

Fig. 4.

Entamoeba invadens nuclear protein(s) specifically bind the *E. invadens* Core Promoter Motif - GAAC-Like (*EiCPM-GL*) and *E. invadens* 3'- U-Rich Motif (*Ei3'-URM*). (A) Electrophoretic mobility shift assays (EMSAs) showed specific binding of *E. invadens* nuclear protein to the *EiCPM-GL*. Competition assays using 2X or 10X excess cold probe for wild type (WT) *EiCPM-GL* demonstrated the specificity of this interaction. Mutated *EiCPM-GL* (Mutant) was not able to compete with the labeled *EiCPM-GL* at either 2X or 10X. (B) EMSAs showed specific binding of *E. invadens* nuclear protein to the *Ei3'-URM*. Competition assays using a 100X excess WT *Ei3'-URM* cold probe demonstrated the specificity of this interaction. Mutated *Ei3'-URM* motif (Mutant) did not compete with the labeled *Ei3'-URM* probe at either 10X or 100X.

Fig. 5. Validation of *Entamoeba invadens* RNA-Seq data by luciferase reporter assay. (A) Expression level (Fragments Per Kilobase of exon per Million mapped reads (FPKM)) of three transcripts - EIN_057870 (T1), EIN_060140 (T2) and EIN_060150 (T3) - throughout

the developmental time points trophozoite (Troph), 8 h, 24 h, 48 h and 72 h encystation, and 2 h and 8 h excystation). All three transcripts showed stage-specific expression with significant expression only in trophozoites. Arrows show the time points selected for luciferase assay (trophozoite and 18 h encystation). (B) Promoter activity of the three constructs, T1, T2 and T3, was determined in stable transformants maintained at 40 µg/ml of G418. Luciferase activity was measured in trophozoites (Troph) and after 18 h of encystation (Cyst). As expected, the promoter-less (-) construct showed no activity. Luciferase assay shows all three promoters are highly active as represented by Relative Luciferase Units (RLUs) in trophozoites. However, after 18 h of encystation all three constructs had low expression. (C) Expression level (FPKM) of four transcripts - EIN_166430 (C1), EIN_147320 (C2), EIN_053400 (C3) and EIN_099680 (C4) - throughout the developmental time points (trophozoite, 8 h, 24 h, 48 h, and 72 h encystation and 2 h and 8 h excystation). All four transcripts showed stage-specific expression with high expression during encystation. Arrows show the time points selected for luciferase assay (trophozoite and 18 h encystation). (D) Promoter activity of four cyst-specific promoters was measured in stable transformants in trophozoites and after 18 h of encystation. Luciferase activity was compared with the minus promoter construct and between trophozoites and cysts. Luciferase assay shows all four promoters demonstrated higher activity at 18 h of encystation than in trophozoites. (E) Expression level (FPKM) of two transcripts - EIN_212890 (Ex1), EIN_239520 (Ex2) - throughout the developmental time points (trophozoite, 8 h, 24 h, 48 h and 72 h encystation, and 2 h and 8 h excystation) is shown. Both transcripts showed stage-specific expression and high expression during the excystation stages. Arrows show the time points selected for luciferase assay (trophozoite and 8 h excystation). (F) Promoter activity of two excystation-specific constructs (Ex1, Ex2) was measured in stable transformants in trophozoites and after 8 h of excystation. The excystation specific promoters (Ex1, Ex2) showed low activity in trophozoites but high expression during excystation. Luciferase assay shows that both promoters have higher activity in excystation samples compared with trophozoites. All experiments were done in triplicate; average and S.D. are shown. * $P < 0.05$, ** $P < 0.005$, *** $P < 0.0005$.

Table 1

Sequences of primers used in this study.

Primers	Sequence (5'-3')
EIN_057870 Sense	ACTAGTCTCGTTTTATTGTCCCG
EIN_057870 Anti-sense	GCTAGCTTCAATATTGATTGACATGTT
EIN_060150 Sense	TCTAGACTCTGGGCTTTGAGTTACA
EIN_060150 Anti-sense	GCTAGCAATTTGAAACAACGACATCTG
EIN_060140 Sense	TCTAGACTCCGTGGACTTTTCCC
EIN_060140 Anti-sense	GCTAGCTATAGCGAAAAGGGACATC
EIN_166430 Sense	ACTAGTCTATTTGTTTGAGTAAAAG
EIN_166430 Anti-sense	GCTAGCACTCACCCGTAAAGAC
EIN_147320 Sense	ACTAGTACGTGCTGGTCTTACC
EIN_147320 Anti-sense	GCTAGCGCCAAAACGTGAAGGC
EIN_053400 Sense	ACTAGTCTCCGTTGTAATATGACGGGT
EIN_053400 Anti-sense	GCTAGCCTTCTTGCTTTGTCCATTTAC
EIN_099680 Sense	ACTAGTCGTGATTGACGTCATTCGGG
EIN_099680 Anti-sense	GCTAGCTCCCATTGCGTAGCTCATTTTC
EIN_239520 Sense	TCTAGAAAATGGATAAGAATTATCCTC
EIN_239520 Anti-sense	GCTAGCCACTGCAATTGTCTCCATC
EIN_212890 Sense	ACTAGTGAAAATAACGGTGCGAC
EIN_212890 Anti-sense	GCTAGCCAGACTATTTCATTTCATT