Mol Genet Genomics (2008) 280:565–574
DOI 10.1007/s00438-008-0387-5

ORIGINAL PAPER

mRNA processing in *Antonospora locustae* spores

Nicolas Corradi · Lena Burri · Patrick J. Keeling

Received: 13 August 2008 / Accepted: 14 September 2008 / Published online: 26 September 2008

Abstract Microsporidia are a group of intracellular parasites characterized by highly reduced and compact genomes. The presence of a high gene density had several consequences for microsporidian genomes, including a high frequency of overlap between transcripts of adjacent genes. This phenomenon is apparently widespread in microsporidia, and strongly correlated with gene density. However, all analyses to date have focused on one or a few transcripts from many loci, so it is unclear how diverse the pool of transcripts at a given locus may be. To address this question, we characterized initiation and termination points from 62 transcripts in gene-dense regions in *Antonospora locustae* spores using both conventional and fluorescence-based RACE-PCR procedures. In parallel, we investigated the abundance and nature of transcripts along a 6 kb region surrounding the actin locus of *A. locustae* using northern blotting, RACE-PCR and previously characterised EST sequences. Overall, we confirmed previous suggestions that most transcripts in *A. locustae* spores overlap with the downstream gene, but that at the 5′ end untranslated regions are very short and overlap is rare. From fluorescence-based RACE-PCR we show that transcription of most genes (31 out of 34) initiates at a single position, whereas 35% of loci analyzed with 3′ RACE polyadenylate mRNA at several sites. Finally, we identified the presence of previously unsuspected and very large transcripts in *A. locustae* spores. Those transcripts were found to overlap up to four open reading frames in different strands, adding a novel layer of complexity in the mRNA transcription of this microsporidian species.

Keywords Overlapping transcription · *Antonospora locustae* · Microsporidia · Genome compaction · Processing points · Northern blotting

Communicated by S. Hohmann.

N. Corradi (✉) · L. Burri · P. J. Keeling
Canadian Institute for Advanced Research,
Department of Botany, University of British Columbia,
3529-6270, University Boulevard, Vancouver,
BC V6T 1Z4, Canada
e-mail: ncorradi@interchange.ubc.ca

Introduction

Microsporidia are a diverse group of obligate intracellular eukaryotic parasites which are common in arthropods, fish and mammals, including humans (Becnel and Andreadis 1999; Larsson 1999). All recent genomic analyses strongly suggest that they are specialized and derived relatives of fungi (James et al. 2006; Keeling 2003; Keeling and Doolittle 1996; Keeling et al. 2000; Van de Peer et al. 2000). Microsporidian nuclear genomes include the smallest known: at the extreme the *Encephalitozoon intestinalis* genome is only 2.3 Mb, and many other species have genomes less than 10 Mb (Biderre et al. 1994; Katinka et al. 2001; Peyretaillade et al. 1998). Genome reduction has taken place through the loss of many genes involved in important metabolic pathways, making them highly dependent on the host for energy production (Goldberg et al. 2008; Tsaousis et al. 2008; Vivares and Metenier 2000), as well as the compaction of the remaining genes. Compaction has in turn taken place through the reduction in the length of the genes, as well as the shortening of intergenic regions. In general microsporidian genomes are characterized by highly divergent gene sequences (Thomarat et al. 2004), but highly conserved gene order (Corradi et al. 2007; Slamovits et al. 2004).

Intergenic spaces are functionally important in many ways. They act as buffers between genes during expression, and also hold regulatory information, such as promoters, enhancers, terminators and polyadenylation sites (Mignone et al. 2002). Some of these elements are not included in transcripts, but many are. In most eukaryotes studied so far, the 5′ untranslated regions (UTRs) of transcripts can be a hundred or more nucleotides in length, while 3′ UTR can be several kilobases (Mignone et al. 2002). Because eukaryotic genomes are usually not very compact, even long UTRs seldom overlap with those of transcripts from adjacent genes (Gerads and Ernst 1998; Hansen et al. 1998), and when they do overlap they can affect the expression of one or both of the affected genes (Prescott and Proudfoot 2002). In highly compacted genomes, such as those of microsporidia and nucleomorphs, the situation seems to be quite different. Specifically, transcripts from adjacent genes have been shown to overlap to a great extent (Corradi et al. 2008; Gilson and McFadden 2002; Slamovits and Keeling 2004; Williams et al. 2005), and to correlate with the length of intergenic region. In microsporidia, the size of UTRs has also been found to vary from a few base pairs to several hundred (Corradi et al. 2008; Slamovits and Keeling 2004). There is no correlation between the size or overlap of UTRs at homologous genes in different species, suggesting this process is quite fluid over time (Corradi et al. 2008).

A number of questions remain to be addressed regarding the frequency and variability of transcript overlap and, most importantly, how the process is controlled and whether they are viable transcripts. To date, the problem has been addressed by sequencing of expressed sequence tags (EST), and sequencing of RACE-PCR products (Corradi et al. 2008; Williams et al. 2005). These techniques give a fair overview of the process, but they do not necessarily allow for the detection of multiple transcripts, nor do they provide information on relative transcript abundance, or indeed any detailed information about transcript variability at any one locus. Here we apply two additional approaches to detect rare transcripts in the microsporidian *Antonospora locustae*. First, we have developed a FAM (5-Carboxyfluorescein)-labeled RACE-PCR procedure that unambiguously discriminates between different transcript endpoints of similar size for a given gene. The 3′ UTR endpoints were also identified for a subset of those genes using conventional 3′ RACE PCR and ESTs, which confirmed the previously observed trends for the species, as well as the reliability of the fluorescence-based methodology. In parallel, the transcript diversity has been investigated in detail in one region of the genome, the genes surrounding the *A. locustae* actin locus. To do this, we first investigated the frequency of overlapping transcription using RACE and EST fragments, and using this data then designed probes for Northern hybridization. Between the two methods a complex pattern of transcription around this locus was described. Overall, these results demonstrate that several independent mechanisms govern mRNA transcription in *A. locustae* spores, and that overlapping transcription accounts for a significant portion of its transcriptome.

Materials and methods

RNA isolation and RACE-PCR procedure

Antonospora locustae spores (M&R Durango, Bayfield, CO) were disrupted by beating with glass beads, and total RNA was extracted using an RNAqueous kit (Ambion, Austin Texas) according to the manufacturer's intructions. We have chosen to use spores over intracellular stages to allow a direct comparison with previous data obtained by others (Corradi et al. 2008; Slamovits and Keeling 2004; Williams et al. 2005). Some of the loci used in this study were identified in the *A. locustae* genome database (http://gmod.mbl.edu/perl/site/antonospora, *A. locustae* Genome Project, Marine Biological Laboratory at Woods Hole, funded by NSF award number 0135272).

3′- and 5′- RACE was performed using the First Choice RLM-RACE kit (Ambion, Austin, TX). For 5′-RACE, this protocol involves sequential treatments with calf intestinal phosphatase (CIP) and tobacco acid pyrophosphatase (TAP) prior to an RNA-RNA adaptor ligation with T4-RNA ligase. This selects for full-length, capped mRNAs. 5′ UTRs were amplified using gene-specific nested primers (see Table 1). 5′-RACE was performed on 43 loci using FAM-labelled RACE-inner primers (5′-CGC GGA TCC GAA CAC TGC GTT TGC TGG CTT TGA TG-3′), and fragments separated in a polyacrilamide matrix using an ABI31700 DNA sequencer (Applied Biosystems). The fluorescent peaks were scored using Peak Scanner™ (Applied Biosystems) and their size annotated. When using a conventional RACE-PCR, a total number of 9 5′ UTRs and 20 3′ UTRS from 12 loci were identified and sequenced. In this case both 5′- and 3′-RACE fragments were ideally analysed for a given locus, but in some cases only one could be amplified using various conditions and primers. When more than a single band was amplified through the PCR reaction, all products were systematically cloned and sequenced to ensure they are derived from the expected region of the genome. In cases where they do, their location and length were included in the figures and their DNA sequence submitted to GenBank. In a few cases, a product was sequenced but discarded, for example 3′ RACE sequences where the poly-A "tail" was found to correspond to an A-track in the genome. In total, 47 loci were analyzed in this study and the 33 cDNA fragments corresponding to

Table 1 List of the *Antonospora locustae* genes analyzed in this study using 5′ RACE PCR and their best match against the *Encephalitozoon cuniculi* genome

Locus	Putative protein function	Outer-primer used	Inner-primer used	5′ length of intergenic region	UTR 1 (bp)	UTR 2 (bp)
AL03_0180	Chromobox protein	CGG ATG CCT CAT ACT CTG CTA TTA G	GCT TGC TCG AAT ATC CTT CCC	856 bp	185	
AL03_0170	Hypothetical protein	TCA AGA TGC TCG CAT TCA TTC TCT A	ACA CCT GGA ATT CAC GCA GTA	571 bp	565	
AL03_0160	Hypothetical protein	GAC TCA TTA GCG AAC ACC TCC TTG T	AAC ATT CTT CAC CTC AAC AAG	95 bp	NA	
AL03_0305	Vacuolar ATP synthase subunit F	ATC TCC TTG CTC ACT GAT ATG AAG T	GAC AAT GGA ATT CCC GTC AGT	113 bp	1	
AL03_0310	40S ribosomal protein S1	CTT CGA GGC CAT CAT GTC GTC T	CGT TGT TCA CTG TAA CGC TGA	21 bp	NA	
AL03_0320	60S ribosomal protein L13	CTT CTC AAC CGG CAT GGG ATA CAT	TTCCGGAGCTGAATTCGCTTCT	21 bp	3	
AL03_0420	Similarity to yeast CDC68 protein	GAC TGG TAA CCG CAT ACG TAT CTG T	GTC TCT GGA AGC TCG TAT CCA	391 bp	NA	
AL03_0430	WD repeats-containing protein	CGT GTC CAG GAC TTC CTT CAG AGT	GCG AGC GAA AGA AGG TGG TTT	327 bp	NA	
AL03_0520	Heat shock related 70 kDa protein	TGA AGT ATG CAG CAC ACG AGT ATG T	TTG CAT CTA TCT CCT TGC TGG	191 bp	35	
AL03_0530	Hypothetical protein	CATACGAATAACCTCCTTGATTGCA	AAC TGG GTG CCT TTC AGA TTG	191 bp	NA	
AL11_1460	Translation elongation factor 2	GCA ATC TTG GCT TTG ACA ACA AGT G	CATGGTCCACATGTGCTATAACAGA	89 bp	NA	
AL11_1450	Transport protein SEC13	ATT GCC TTT GTC ACC GGT CCT GGA	GAA GAG GCT GTT ATG ATG CGT	NA	NA	
AL11_1380	60S ribosomal protein L15	CTT ATG CGC ACT CTG AAG ATG CAT	AGC AGT GTT GAG CCT GAA CTC	108 bp	3	
AL11_1390	Hypothetical protein	GAT TGA AAA CGT CTG CGT GCA CTG	CGC ATA ACA TAG GAA TGT ACG CAC A	108 bp	1	28
AL11_0660	Ser/Thr protein phosphate PPI-1 catalytic S	GAAGGTCCTGATACTGTCCGTGTAT	TC GCA TAG GTA CCT TAT CTC CTG T	389 bp	9	
AL11_0670	Hypothetical protein YG22	TAT CTG CAG CTG TTT GCA ATC TTC A	CTG TGT TGC TGC TGT ATG CAT	510 bp	2	
AL02_1100	Heat-shock protein HSP90	TTA TCG CAT GCA TCG CTT GCA TT	ACG TAT GAA GAA CTC CTT AGA	235 bp	9	
AL02_1090	ATP-dependent DNA-binding helicase	TAT GCA AGG GCT GCA CAC AGA ATG	TTC CTC CCT CTT GCA GTG CCA	313 bp	NA	
AL07_0130	Anti-silencing protein 1	GCG TCT CCA GAG TAT ATA ACA TCG A	TTT GTG CAC TCA AGT ACA ACA	211 bp	5	
AL07_0120	Hypothetical protein	TGG CTC GAG AGA AAG AGT CAG ATA T	AG TGA CTT GCA GAT GGG ACA TCT A	78 bp	43	
AL07_0190	26S proteasome regulatory subunit 10	TTATCCTCGATTTGCAGATACTCCA	GCT TGC ACT TTA CAG CTG GAT	208 bp	0	
AL07_0200	Hypothetical protein	CTG GAT TTC ATG ACA ACT CTG CCT A	TCT TTG ACA GAT GCT CGT CAA	>1 kb	22	
AL07_0210	Hypothetical protein	CAA CCA TCG ACA GAG ACT GGA CAT	AGA ATG TTG TGC TGC TCT TTC	103 bp	98	
AL07_0260	Similarity with WD-repeat proteins	GCA AGC CAC ACT TGC CCA AAC TT	GGC CTT CCA AGA TCA AAG TCG	12 bp	0	
AL07_0270	Hypothetical protein	AAT AGT AGC GCG TAG ACA GAT GGA	ACA AAT GCC GGT GTT GAG CGT	151 bp	0	26
AL07_1260	Guanosine diphosphatase	CCA GTG ATA GGA GTG CAC AGT AGA A	TGG TAC CTG TTG ATC CTG CGT	223 bp	3	
AL07_1250	Hypothetical protein	CCG GTT CTA TTC TCT CCA GGT AGT A	GTA CGC CTT CAT CAT GCA CCT	223 bp	22	
AL08_1220	Hypothetical protein	CTA GGC AGC GAA GAA GGT CGT AT	CTT TGT TGT GTT GGA GAC GTC	639 bp	4	
AL08_1210	Hypothetical protein	ACA CTA GGT ACT GCG GAA CGT TCA	TTT CGT ACA ACA AGC TCC TCC	764 bp	1	

Table 1 continued

Locus	Putative protein function	Outer-primer used	Inner-primer used	5′ length of intergenic region	UTR 1 (bp)	UTR 2 (bp)
AL08_1100	Coatomer complex beta subunit	GTC CAC GTA GCT TGT TCC ATC GAT	GCA CAA TCA CTT GTT CAA TGC	564 bp	10	160
AL08_1110	Guanine nucleotide binding protein beta SU	GAC CTT GCA ACC GCC ACA TCG TT	ACTCCTTCACGATTCTGCCAAACAT	97 bp	**425**	
AL05_1240	Gamma glutamyl transpetidase	GCT CCA GCG TTC TCA CTT GTA CT	AAA TCT TTA GTC TCT TGA TGG	459 bp	**803**	
AL05_1250	CDP-diacylgycerol synthase	CTTGGGACATTTTCATTGCATAGCT	TGCAAGTTCTGCTTAGCTTTCATCA	>1 kb	11	
AL04_1670	Hypothetical protein	TTT TAG AGC AAT GCC AAT GTT TGG A	CGC CTG TTC CAA CTT GCT ACT	248 bp	10	
AL04_1660	Hypothetical protein	TTG AAA CCC GTA GTT CGT GTT GAA GTT	TT GAT AGA CTC GGG CTC CAC ATG T	165 bp	14	
AL03_1420	Hypothetical protein	CAA TCA CAA GTA TCA GGA GCA GCT G	GCT GTT CTG AAA GAG AAG CTG	67 bp	0	
AL03_1430	RAS-Like GTP-binding protein PROTEIN YPT	TCC TCG CCG TCA ACT GTA ATT GTC T	CTA CGC CAA TCG TGC TTA TGT	69 bp	NA	
AL09_0190	Putative leucine repeat-rich protein	GAC ATG CTC AAA TGT AAA GCA GGT A	GTC CAT ATG CCT AGA ATG CAC	99 bp	54	
AL09_0180	Hypothetical protein	AAG CTG CTG CCC TAA ACG AAG TA	GTA TTG TAG GCA CTT AGC ACA	199 bp	8	
AL09_0170	GTP-binding protein	TTCTTTATCGTCCACATCCACCTTC	ATT CCT ATT GTC GGT TCG TAG	199 bp	1	
AL09_0160	Hypothetical protein	AGC CAT TTA AGA GCC CGG TCT TGT	AAA CCT GCT TCA CGA ACG CTG	43 bp	**661**	
AL07_1620	BOS1-like vescicular transport protein	CGT GCT GTA GGA CGT TAT TAG AAT G	TAG ATG ATG GAG CAT TTG GAG	317 bp	4	
AL07_1630	Putative protein with mut T domain	ATC ATG GAA CGA GAG ATA CAG CGA	GAT AAT GAA GTA AAG GCG CTC	84 bp	47	

A. locustae loci are designated according to the designation of the *E. cuniculi* ORF (i.e. AL07_1630 from *A. locustae* correspond to the orthologue of ECU07_1630 from *E. cuniculi*). Their putative function of the analyzed loci, the length of their 5′ untranslated regions (UTRs), the length of the 5′ intergenic regions, the primers we used in the RACE PCR reaction, and the length of the UTRs we identified are also listed. The UTRs showing overlap with upstream genes are shown in bold. *UTR* untranslated transcribed region

5′ and 3′ transcriptional ends have been submitted to GenBank under accession numbers FK250482–FK250514.

Northern blot analysis

Total RNA was extracted from *A. locustae* spores (M&R Durango, Bayfield, CO) by bead beating followed by RNA isolation using the RNAquous kit from AMBION (Austin, TX). 6 μg of denatured RNA was electrophoresed in a 1.2% formaldehyde agarose gel and transferred onto a Hybond-N+ membrane (GE Healthcare, Piscataway, NJ). Northern hybridization was carried out using the Alkphos Direct Labeling and Detection System from GE Healthcare according to the manufacturer's protocol. A DNA probe representing 605 bp of the *A. locustae* actin sequence labeled with alkaline phosphatase was used for hybridisation and CDP-star as detection reagent.

Results

Identification of 5′ UTRs in *A. locustae* using FAM-labelled oligos

A total of 43 loci with clearly recognisable upstream genes were identified from the *A. locustae* genome database (http://gmod.mbl.edu/perl/site/antonospora). Using 5′-RACE with FAM-labelled inner-primers, the transcription initiation sites were determined for all genes but nine, for which RACE products could not be obtained (Table 1, Fig. 1). Of the remainder, a transcript from one gene encoding a hypothetical Protein (AL06_0160) was found to initiate a few base pairs beyond an upstream gene, and transcripts from two other genes, encoding a guanine nucleotide binding protein (AL08_1100) and a gamma glutamyl transpeptidase (AL05_1240), were found to start within upstream genes. Transcripts from the remaining 31 genes initiated in the upstream intergenic region, in most cases immediately upstream of the start codon. This of course does not rule out the possibility that transcripts from these genes overlap with transcripts from the upstream gene in cases where the 3′ UTR of the upstream gene extends into the region identified here as the 5′ UTR of the downstream gene. Similarly, in only three cases were multiple initiation sites observed: for two hypothetical proteins (AL11_1390 and AL07_0270), and a gene encoding a coatomer complex beta subunit (AL_1100). The length of 5′ UTRs varied considerably from gene to gene. The shortest corresponded to capped AUG codons, while the largest was 803 bp long and started in the middle of an upstream gene, a feature already reported for the photolyase from this species (Slamovits and Keeling 2004). The average length of *A. locustae*

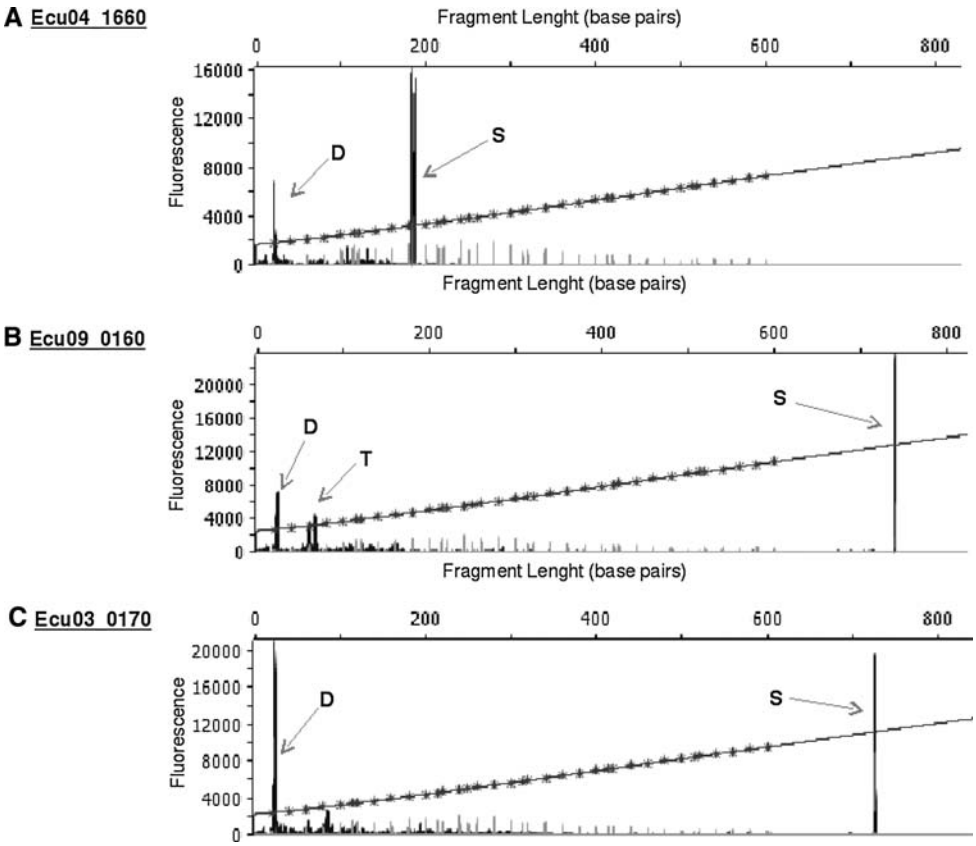

Fig. 1 Examples of 5′ RACE PCR fragments analysis using FAM-labeled oligos and analyzed using the Peak Scanner™ Software v1.0 (Applied Biosystems). The horizontal axis represents the length of the fragments in base pair. The vertical axis represents the intensity of the fluorescent probe. To ensure a conservative analysis of the 5′ RACE products, only peaks indicated by a "S" have been used in the subsequent analyses of overlapping transcription. All other peaks have been discarded because they were corresponding to dimer-primers (D) or because they showed signal intensity lower than half of the biggest peak. Other peaks were discarded from subsequent analyzes despite their strong fluorescence (T), because they may have equally reflected a truncated 5′ RACE product for our gene of interest or, as previously identified (Corradi and others 2008), the start of a overlapping 5′ UTR of downstream genes

5′ UTRs was 41 bp (Table 1). Overall, most genes seem to have a single, clear and well-defined initiation point within the upstream intergenic space.

Reliability of the FAM-labelled RACE-PCR and overall patterns of transcription at the eight selected loci

The reliability of the FAM-labelled RACE methodology was tested by selecting eight loci with a range of 5′ UTR length. RACE-PCR was reproduced for those genes using conventional primers and the products characterised by cloning and sequencing. The sequence of the selected PCR products always corresponded to the targeted genomic region, and their length corresponded to that determined by analysis of FAM-labelled products. A representation of these 5′ UTRs is shown in Fig. 2.

The loci represented in Fig. 2 represent a number of other interesting features, such as converging or diverging genes, and variable lengths of intergenic regions. Therefore, we further examined the variation in processing points in these regions using 3′ RACE and publicly available EST data. The RACE procedure identified several 3′ UTRs, with variable lengths, most of which overlapped with downstream genes. We also identified several cases where more than a single poly-A tail could be isolated. Interestingly, a few of those UTRs were found to harbor potential consensus polyadenylation motifs (AAUAAA), located upstream of the Poly-A tail. However, these motifs were rare among the 3′ UTRs we identified, and their presence was not linked with either presence or absence of transcription overlap.

Transcription of the actin gene and close neighbors in *A. locustae* spores

The methods used to date to examine the ends of microsporidian spore transcripts are informative for identifying processing points, but it is impossible to link the 5′ and 3′ data generated by these methods to completely determine the extent of any given transcript. To provide a more thorough picture for one locus and its surrounding genes, we combined data from cDNA ends with Northern blot data on total transcript length. Total mRNA from *A. locustae* spores was probed using a 605 bp region of the *A. locustae* actin gene (Fig. 3). This probe recognised two major transcripts of 1.6 and 3 kb, but three other distinct transcripts were also observed, ranging in size from 1.4 to over 2 kb. To determine where these transcripts lay with respect to the actin gene, we identified all open reading frames in approximately 6 kb of the genome centred on the actin gene. Nine ORFs, four of which shared high sequence similarity with *E. cuniculi* genes were identified in this region, as well as a single putative polyadenylation motif. 5′ and 3′ RACE was

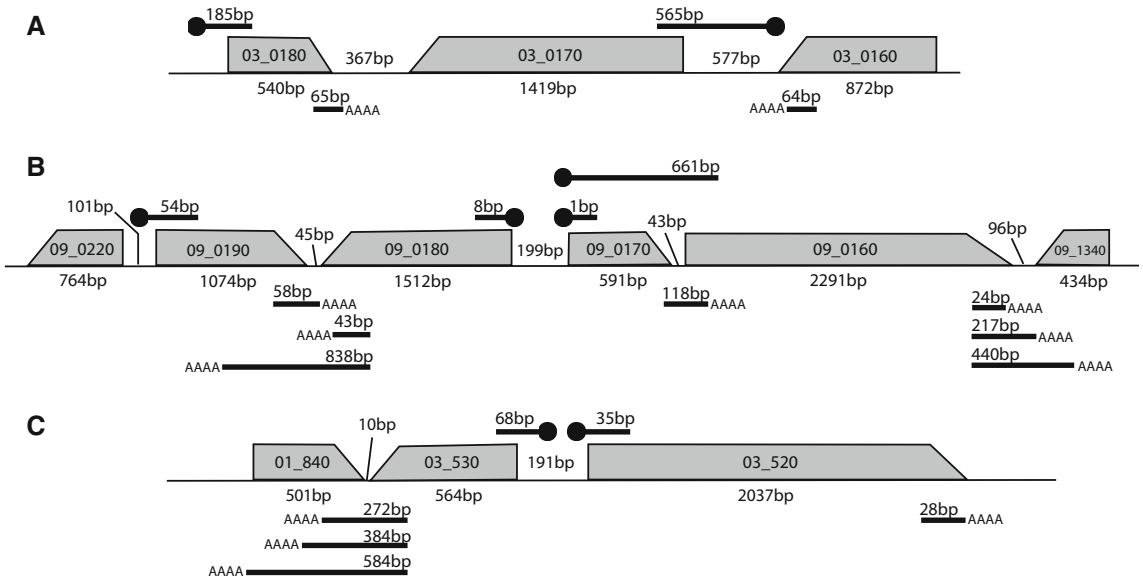

Fig. 2 Patterns of transcription identified among genomic regions in *Antonospora locustae*. *Grey rectangles* represent the position of the genes in genomic DNA (with gene names inside) and the angle of the rectangle represents its transcriptional direction. Length and position of the 3′-RACE fragments are represented by *black rectangles* and their respective poly-A tail. Length and position of the 5′-RACE fragments are represented by *black rectangles*, with the cap represented by a *dot*. The blunt part of the rectangles represents the location of the primer we used to perform the PCR reaction. The length in base pairs (bp) of the genes and the intergenic regions is shown. The position of the *A. locustae* EST fragments (Williams et al. 2005) are represented by the white rectangles and their respective poly A tail. The length of the 3′ and 5′ UTRs, as well as the EST fragments is shown above the rectangles representing these fragments (bp). The location of polyadenylation motifs (AAUAAA) is shown (M)

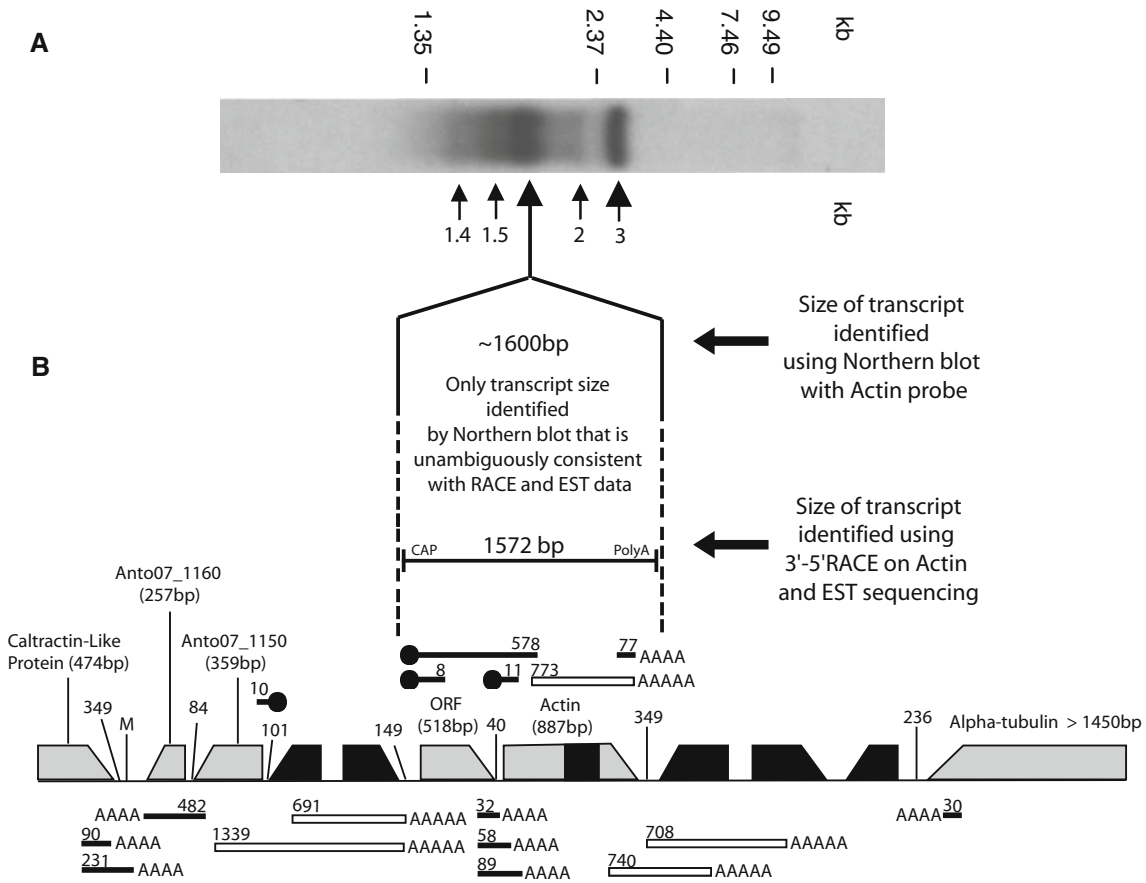

Fig. 3 **a** size of mRNAs and their level of expression identified by northern blotting using a DIG-labeled probe encompassing 605 bp of the actin gene. Sizes of the different transcripts are indicated. The only transcript identified by Northern that has a size consistent with transcripts identified by RACE and EST sequencing is highlighted. **b** Overall pattern of transcription along a 6,922 bp genomic region in *Antonospora locustae* and identified by RACE-PCR and EST sequencing. *Grey* and *angled rectangles* represent the position and direction of genes showing significant sequence similarity with genes from *Encephalitozoon cuniculi* and with the angle representing their transcriptional direction. 1. Caltractin-Like Protein (474 bp); 2. Anto07_1160 (257 bp); 3. Anto07_1150 (359 bp); 4. Hypothetical Protein (518 bp); 5. Actin (887 bp); 6. Alpha-Tubulin (incomplete, >1,403 bp). *Black* and *angled rectangles* represent the position of putative ORFs with no homologies with sequences deposited in public databases and from which no RACE-PCR product could be obtained. The position and length of the probe designed on the actin locus is represented by the *black rectangle*. Length and position of the 3′-RACE fragments are represented by *black rectangles* and their respective poly-a tail. Length and position of the 5′-RACE fragments are represented by *black rectangles*, with the cap represented by a *dot*. The blunt part of the rectangles represents the location of the primer we used to perform the PCR reaction. The length in nucleotides (bp) of the genes and the intergenic regions is shown. The length of the 3′ and 5′ UTRs is shown above the rectangles representing the RACE-PCR products. The total length and position of the *A. locustae* EST fragments (Williams et al. 2005) are represented by the *white rectangles* and their respective poly-A tail. The location of a polyadenylation motif (AAUAAA) is shown (M)

performed on the ORFs surrounding the actin locus, and all EST fragments corresponding to this region of the genome, and characterized by the presence of a Poly-A tail, were identified in public databases (Williams et al. 2005). This analysis resulted in the identification of 11 terminators and 3 initiators along the 6 kb long genomic region. Gathering this data together on the map of this region of the genome produces a detailed picture of the transcriptional mechanisms occurring in *A. locustae* spores (Fig. 3). This picture is consistent with what has been identified for other loci in this study (Fig. 2) and in previous reports (Corradi et al. 2008; Slamovits and Keeling 2004; Williams et al. 2005), but shows the presence of a previously unsuspected diversity in transcripts. The 1.6 kb major transcript corresponds in size to a transcript beginning at the major initiation site identified upstream of ORF 4 and ending at the major termination site downstream of actin. In contrast, there is no band on the Northern blot that corresponds to the actin transcript on its own. The major 3 kb transcript cannot easily be assigned to any particular location as it is so large.

Discussion

Contrasting transcription initiation and termination in *A. locustae*

Distinctive trends in how transcription is controlled at both the 5′ and 3′ ends have emerged from analysis of many *A. locustae* genes. For example, from 40 initiation points identified in this study, only 3 (7.5%) were located within or beyond the upstream gene (Table 1, Fig. 2). In contrast, more than 50% of the 3′ UTRs (Figs. 2 and 3) extend well into or beyond downstream genes. Similarly, only 8% of the genes have transcripts that initiate at more than one position, whereas 35% have multiple polyadenylation sites. The length of 3′ and 5′ UTRs seem to correlate differently with the length of intergenic regions: there is no clear link between transcriptional overlap and short intergenic spaces at the 5′ end, but there does appear to be such a correlation at the 3′ end. Specifically, 100% of the genes harboring intergenic regions shorter than 150 bp had 3′ UTRs overlapping with downstream genes, as opposed to 18% of 5′ UTRs.

Overall, the short intergenic regions in *A. locustae* have affected initiation and termination of transcription differently. Initiation is almost always at a single position within intergenic regions, and close to the initiation codon, all features consistent with tight control. In contrast, polyadenylation often occurs at several different positions, far from the end of a gene and within adjacent genes, all suggestive of a more relaxed system. Canonical polyadenylation sites can be identified in the *A. locustae* genome, but they are not common and apparently are frequently unused.

Is the correlation between intergenic regions and transcription overlap functionally relevant in *A. locustae*?

Current and previous reports (Corradi et al. 2008; Gilson et al. 1997; Gilson and McFadden 2002; Williams and others 2005) have demonstrated that in the most compacted nuclear genomes, the probability of overlapping transcription is correlated with the length of intergenic regions, especially in the 3′ regions of genes. However, it is still unclear whether this is a trivial correlation (e.g., transcripts of equal length are more likely to read into another gene if the intergenic region is short) or one with functional significance (e.g., short intergenic regions contain too little information to support termination). If the latter is true, then 3′ UTRs for genes with short 3′ intergenic distances should be longer than average. Analysing the 21 cases now available for *A. locustae*, this may be the case: genes with short 3′ intergenic distances have, overall, slightly longer UTRs. However, the correlation is rather weak and not conclusive evidence that the reduction of intergenic spaces alone led to reduced transcription control. Sampling of 3′ UTRs from many hundreds of genes might shed more light on this question, but at present the data are consistent with the possibility that the shrinking genome led to changes in the way transcription termination is controlled. At the 5′ end there is no such correlation: the size of 5′ UTRs encompassing short intergenic regions were neither significantly larger, nor smaller when compared to the overall UTRs identified.

Do *A. locustae* spore transcripts contain operons?

The actin locus of *A. locustae* presents an interesting problem. Actin is abundant in the *A. locustae* ESTs, and these cDNAs consistently end at a single position immediately downstream of actin. Using 3′ RACE from actin identifies the same position, while 3′ RACE from the upstream unidentified ORF identifies several polyadenylation sites within actin (Fig. 3). Interestingly, 5′ RACE from within actin consistently identified an initiation site at the 5′ end of the upstream unidentified ORF (Corradi et al. 2008) and only and apparently rare initiator upstream of actin (Fig. 3). Northern blotting is consistent with RACE and EST data, since the most abundant RNA we identified corresponds exactly with the size expected for a transcript that extends from the main initiation point upstream of the ORF to the main polyadenylation point downstream of actin (Fig. 3).

A similar situation appears to be true for other genes based on RACE data. Specifically, 5′ RACE from genes encoding the ribosomal protein L34 (Williams et al. 2005), and a polyubiquitin (Corradi et al. 2008) have been found to entirely overlap with their respective upstream gene. Thus, although it is clear that most cases of overlapping transcription in microsporidia do not represent co-expression of genes from a single mRNA, it is possible that functional co-transcription also happens in a few special cases.

However, to conclude that proteins are co-expressed from a single mRNA, we need to know that both proteins are actually expressed, and that they are not represented by individual mRNAs. In the case of the *A. locustae* spores, we first must know that mRNA in the spore is functional, and this is not presently known since spores likely are not metabolically active. Assuming the same expression patterns are found in active cells, there are two possibilities. Taking the actin locus as an example, the present data are consistent with a model where actin is not expressed at high enough levels to have been detected by the Northern blotting used here, whereas the upstream ORF is highly expressed but lacks any termination and polyadenylation signals in its unusually short 3′ intergenic region (40 bp), and uses one downstream of the actin gene instead. One might argue that a single mRNA is used for the translation of both the ORF and actin, or other pairs of genes in the

same strand for which a 5′ initiation site for the downstream gene cannot be identified (Williams et al. 2005).

Operon-like co-transcription is known in other nuclear genomes, in particular *Caenorhabditis elegans* and in trypanosomes (Blumenthal 1998; Blumenthal 2004; Blumenthal and Gleason 2003; Gilson et al. 1997; Gilson and McFadden 2002; Spieth et al. 1993; Zorio et al. 1994), where it is always associated with post-transcriptional processing to individual mRNAs. However, these well-described cases also correlate with the presence of a small spliced leader at the 5′ end of mRNAs, which leads to the processing of polycistronic mRNAs. No such mechanism is evident in microsporidia. The Northern blot did reveal several abundant transcripts that could not be linked with any previous observation for this particular region. Specifically, we identified some very large and by necessity multigene transcripts in *A. locustae* spores (the largest being 3 kbp in length and containing up to 4 potential ORFs). The abundance of these RNAs suggests that they are actively expressed. Importantly, however, none of these large RNAs can represent a simple operon-like molecule because many of the ORFs surrounding actin are encoded on opposite strands. These transcripts more likely represent some other process adding another level of complexity to transcription in these genomes.

Concluding remarks

Overall, we identify two independent kinds of mRNA in *A. locustae* spores: single-gene transcripts (canonical mRNAs with the major exception that many include 3′ UTRs that overlap with downstream genes), and very large transcripts encompassing several ORFs, sometimes located on different strands. It is clear that conventional transcriptional mechanisms have been maintained in *A. locustae*, and probably account for most or all of the gene expression. However, even in these relatively simple transcripts there are a large fraction where the 3′ UTR overlaps with downstream genes, and how this does or does not affect the expression of those genes remains an interesting problem. The possibility that a small fraction of genes are expressed on polycistronic messages still needs to be formally excluded, and if this is indeed found to be true, it suggests that mRNAs are processed post-transcriptionally, or microsporidian ribosomes are unusually able to recognise more than one cistron per message (a seemingly unlikely possibility). Lastly, the presence of abundant and very large RNAs where no single gene can be identified as the target for expression adds another layer of complexity to these already complicated and apparently wasteful systems. As a final note, it is important to stress that to date all analyses of transcription in microsporidia have been performed on RNA extracted from spores, where translation is apparently inactive. This leaves open the possibility that none of these transcripts are functional in the sense that the genes encoded on them are ever expressed. The next steps will have to include both the examination of transcription in metabolically active cells, as well as identifying the functional roles of each transcript type in *A. locustae* spores.

Acknowledgments This work was supported by a grant from the Canadian Institutes for Health Research (MOP-84265). PJK is a Fellow of the Canadian Institute for Advanced Research and a Senior Scholar of the Michael Smith Foundation for Health Research. NC was partly supported by a fellowship from the Swiss National Science Foundation (PBLAA - 114238). L.B. was supported by fellowships from MSFHR and CIHR. We thank Todd Harper for critically reading of the manuscript.

References

Becnel JJ, Andreadis TG (1999) Microsporidia in insects. In: Witter M, WLM (ed) The microsporidia and microsporidiosis. American Society of Microbiology Press, pp 447–501

Biderre C, Pages M, Metenier G, David D, Bata J, Prensier G, Vivares CP (1994) On small genomes in eukaryotic organisms: molecular karyotypes of two microsporidian species (Protozoa) parasites of vertebrates. C R Acad Sci III 317(5):399–404

Blumenthal T (1998) Gene clusters and polycistronic transcription in eukaryotes. Bioessays 20(6):480–487

Blumenthal T (2004) Operons in eukaryotes. Brief Funct Genomic Proteomic 3(3):199–211

Blumenthal T, Gleason KS (2003) *Caenorhabditis elegans* operons: form and function. Nat Rev Genet 4(2):112–120

Corradi N, Akiyoshi DE, Morrison HG, Feng X, Weiss LM, Tzipori S, Keeling PJ (2007) Patterns of genome evolution among the microsporidian parasites *Encephalitozoon cuniculi*, *Antonospora locustae* and *Enterocytozoon bieneusi*. PLoS ONE 2(12):e1277

Corradi N, Gangaeva A, Keeling PJ (2008) Comparative profiling of overlapping transcription in the compacted genomes of microsporidia *Antonospora locustae* and *Encephalitozoon cuniculi*. Genomics 91(4):388–393

Gerads M, Ernst JF (1998) Overlapping coding regions and trancriptional units of two essential chromosomal genes (CCT8, TRP1) in the fungal pathogen *Candida albicans*. Nucleic Acids Res 26(22):5061–50616

Gilson PR, Maier UG, McFadden GI (1997) Size isn't everything: lessons in genetic miniaturisation from nucleomorphs. Curr Opin Genet Dev 7(6):800–806

Gilson PR, McFadden GI (2002) Jam packed genomes–a preliminary, comparative analysis of nucleomorphs. Genetica 115(1):13–28

Goldberg AV, Molik S, Tsaousis AD, Neumann K, Kuhnke G, Delbac F, Vivares CP, Hirt RP, Lill R, Embley TM (2008) Localization and functionality of microsporidian iron-sulphur cluster assembly proteins. Nature 452(7187):624–628

Hansen K, Birse CE, Proudfoot NJ (1998) Nascent transcription from the nmt1 and nmt2 genes of *Schizosaccharomyces pombe* overlaps neighbouring genes. Embo J 17(11):3066–3077

James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J and others (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature 443(7113):818–822

Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Others (2001)

Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 414(6862):450–453

Keeling PJ (2003) Congruent evidence from alpha-tubulin and beta-tubulin gene phylogenies for a zygomycete origin of microsporidia. Fungal Genet Biol 38(3):298–309

Keeling PJ, Doolittle WF (1996) Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. Mol Biol Evol 13(10):1297–1305

Keeling PJ, Luker MA, Palmer JD (2000) Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. Mol Biol Evol 17(1):23–31

Larsson JIR (1999) Identification of microsporidia. Acta Protozoologica 38(3):161–197

Mignone F, Gissi C, Liuni S, Pesole G (2002) Untranslated regions of mRNAs. Genome Biol 3(3):REVIEWS0004

Peyretaillade E, Biderre C, Peyret P, Duffieux F, Metenier G, Gouy M, Michot B, Vivares CP (1998) Microsporidian encephalitozoon cuniculi, a unicellular eukaryote with an unusual chromosomal dispersion of ribosomal genes and a LSU rRNA reduced to the universal core. Nucleic Acids Res 26(15):3513–3520

Prescott EM, Proudfoot NJ (2002) Transcriptional collision between convergent genes in budding yeast. Proc Natl Acad Sci USA 99(13):8796–8801

Slamovits CH, Fast NM, Law JS, Keeling PJ (2004) Genome compaction and stability in microsporidian intracellular parasites. Curr Biol 14(10):891–896

Slamovits CH, Keeling PJ (2004) Class II photolyase in a microsporidian intracellular parasite. J Mol Biol 341(3):713–721

Spieth J, Brooke G, Kuersten S, Lea K, Blumenthal T (1993) Operons in *C. elegans* polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. Cell 73(3):521–532

Thomarat F, Vivares CP, Gouy M (2004) Phylogenetic analysis of the complete genome sequence of *Encephalitozoon cuniculi* supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes. J Mol Evol 59(6):780–791

Tsaousis AD, Kunji ER, Goldberg AV, Lucocq JM, Hirt RP, Embley TM (2008) A novel route for ATP acquisition by the remnant mitochondria of *Encephalitozoon cuniculi*. Nature

Van de Peer Y, Ben Ali A, Meyer A (2000) Microsporidia: accumulating molecular evidence that a group of amitochondriate and suspectedly primitive eukaryotes are just curious fungi. Gene 246(1/2):1–8

Vivares CP, Metenier G (2000) Towards the minimal eukaryotic parasitic genome. Curr Opin Microbiol 3(5):463–467

Williams BA, Slamovits CH, Patron NJ, Fast NM, Keeling PJ (2005) A high frequency of overlapping gene expression in compacted eukaryotic genomes. Proc Natl Acad Sci USA 102(31):10936–10941

Zorio DA, Cheng NN, Blumenthal T, Spieth J (1994) Operons as a common form of chromosomal organization in *C. elegans*. Nature 372(6503):270–272