ELSEVIER

Molecular & Biochemical Parasitology 122 (2002) 171–180

MOLECULAR & BIOCHEMICAL PARASITOLOGY

www.parasitology-online.com

Full-length-enriched cDNA libraries from *Echinococcus granulosus* contain separate populations of oligo-capped and *trans*-spliced transcripts and a high level of predicted signal peptide sequences

Cecilia Fernández [a,b], William F. Gregory [a], P'ng Loke [a], Rick M. Maizels [a,*]

[a] *Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK*
[b] *Immunology Department, Faculty of Chemistry, Universidad de la República, Montevideo, Uruguay*

Received 11 February 2002; accepted in revised form 29 April 2002

Abstract

The tissue-dwelling larval stages of the cestode *Echinococcus granulosus* are intimately associated with the host, implying that a range of molecular mediators may be secreted by the parasite into the host environment. These mediators are being sought through a transcriptome-based analysis, using recombinant cDNA libraries. Conventional cDNA libraries of *E. granulosus* contain high levels of mitochondrial transcripts, as well as host (bovine) genomic DNA. In particular, 60% of a conventional protoscolex stage cDNA library corresponds to the large subunit (LSU) of mitochondrial rRNA. We attribute the presence of LSU rRNA copies to its polyadenylation in *E. granulosus*. To circumvent this problem, we adapted the 5′ Rapid Amplification of cDNA Ends (RNA-ligase mediated RACE) technique that excludes all polynucleotides missing the 7-methyl-guanosine (7MG) cap specific to the 5′ end of full-length mRNA. By ligating a specific oligonucleotide (oligo-cap) to 7MG-bearing mRNA, three cDNA libraries were made by PCR from oligo-cap and oligo-dT primers. Analysis of these libraries showed that mitochondrial RNA contaminants had been excluded. Moreover, no bovine genomic sequences were detected. In parallel, we constructed three cDNA libraries using the newly described *trans*-spliced leader (SL) from *Echinococcus*. Although these represent a smaller subset of parasite genes, mitochondrial and genomic contributions were again excluded. In both cases, a majority of cDNAs (61–92%) were judged to contain the initiation ATG codon, and 11–27% of inserts included potential N-terminal signal sequences. The 5′ UTR tracts of most oligo-capped cDNAs were <100 nt, although $\sim 8\%$ were longer than this. Among the *trans*-spliced cDNAs, 43% potentially utilise the AUG donated by the SL, and in only 6% was the SL separated from an endogenous putative start site by >60 nt. Sequence analysis of randomly selected clones shows virtually no overlap between the oligo-capped and SL libraries, indicating that *trans*-spliced *E. granulosus* mRNAs appear to be insensitive to the enzymatic treatments used to 'oligo-cap' unspliced mRNAs. The oligo-capped and SL strategies represent efficient and complementary pathways to isolate full-length cDNA clones from this cestode parasite and, possibly, from related parasitic flatworms.

Keywords: Cestode; Genome; Helminth; Open reading frames; Untranslated regions

1. Introduction

Molecular biology has transformed parasitology, with perhaps the most striking examples coming from our ability to sequence and express the protein repertoire of parasites through recombinant cDNA libraries. However, different parasitic organisms vary greatly in the detail of transcriptional mechanisms and levels of RNA processing, and indeed many fundamental observations such as *trans*-splicing [1] and RNA editing [2] have emerged from careful analysis of parasite sequence data.

Abbreviations: CW, hydatid cyst wall (germinal and laminated layers); EST, expressed sequence tag; LSU and SSU, large (16S) and small (12S) subunits of mitochondrial rRNA; 7MG, 7-methyl-guanosine; PS, protoscoleces; RACE, rapid amplification of cDNA ends; SL, spliced leader; TMG, 2,2,7-trimethyl-guanosine; UTR, untranslated region.

* Corresponding author. Tel.: +44-131-650-5511; fax: +44-131-650-5450

E-mail address: rick.maizels@ed.ac.uk (R.M. Maizels).

0166-6851/02/$ - see front matter
PII: S0166-6851(02)00098-1

Cestode parasites are a major group of helminths of medical and veterinary importance [3], but no genomic initiative has yet been adopted for these organisms. We recently launched a gene discovery project on the dog tapeworm *Echinococcus granulosus*, the agent of cystic hydatidosis in humans and domestic and wild animals [4]. Our aim is to identify molecules involved in the host–parasite cross-talk, and for this purpose we have prepared cDNA libraries from tissue-dwelling larval stages. The larva has the form of a fluid-filled hydatid cyst, bounded by the cyst wall (CW). The latter comprises an innermost 'germinal layer' of live parasite tissue, which synthesises an outer carbohydrate-rich 'laminated layer'. The germinal layer also gives origin, through budding towards the interior of the cyst, to the larval worms (protoscoleces (PS)). Protoscoleces are the stage capable of infecting dogs and maturing to adult worms.

Gene discovery through conventional cDNA libraries has certain limitations, such as contamination with rRNA and mitochondrial mRNA transcripts, and with genomic DNA. Moreover, cDNA synthesis using oligo-dT to prime reverse transcription from the 3′ poly-A tail of mRNA [5] does not ensure that all transcripts extend to the 5′ terminus and, as a result, most cDNA libraries contain large proportions of truncated clones. In the case of parasite cDNA libraries, it may also be difficult to avoid contamination with host cDNA or genomic clones.

A convenient strategy, which applies to some parasitic organisms, is to exploit the 5′ *trans*-spliced leader (SL) sequence to isolate full-length cDNAs [6–8]. In trypanosomatids, all mRNAs are *trans*-spliced at the 5′ end with an identical 35-nt oligonucleotide [1,9], while in nematodes a fraction of mRNAs are similarly *trans*-spliced with a 22-nt leader sequence [9–12]. Parasite mRNAs with a 5′ SL can readily be amplified using primers to the SL and 3′ poly-A tail, even from diminutive amounts of parasite material, and despite the presence of host contaminants. Recently, a *trans*-spliced 36-nt leader sequence has been discovered in cestodes of the genus *Echinococcus*, permitting this strategy to be applied to *E. multilocularis* [13], although only a minority of transcripts are included by this method.

In this report, we compare the results obtained using the above methods, with a new approach to construction of cDNA libraries. This simultaneously reduces contamination with transcripts other than nuclear-encoded mRNAs, and ensures a high proportion of full-length cDNAs. The strategy ('oligo-capping') was originally described by Suzuki et al. [14], and recently applied to the large scale preparation of full-length enriched cDNA libraries from humans [15] and the malaria parasite [16]. After ligating a specific oligonucleotide (oligo-cap) to capped mRNAs, reverse transcription is primed with a tagged oligo-dT. This tag is then used for cDNA amplification together with a primer directed against the oligo-cap, and cDNAs are directionally cloned in a plasmid vector. In parallel, we prepared libraries of *trans*-spliced transcripts, which should also contain full-length cDNAs derived exclusively from nuclear-encoded mRNAs. We compare the quality of the libraries, analysing in particular the level of transcripts containing putative signal peptide sequences, as this group is likely to encode the proteins secreted by parasites in order to modulate their host environment.

2. Methods

2.1. E. granulosus parasites and RNA isolation

Hydatid cysts were obtained from the lungs of naturally infected bovines in Uruguay. The hydatid fluid was aseptically aspirated with a vacuum pump. Once settled, PS were recovered from the aspirated fluid and extensively washed in phosphate-buffered saline to remove dead worms and CW debris. They were then observed by light microscopy for viability (flame cell activity and eosin-exclusion). PS batches showing viability in excess of 90% were stored at −80 °C in Trizol reagent (GibcoBRL) until RNA extraction (100 μl of packed PS were routinely kept in 900 μl of Trizol). One fraction of freshly isolated PS was incubated with 0.5 mg ml^{-1} pepsin (in Hanks' solution, pH 2.0, for 3 h at 37 °C), prior to treatment with Trizol. The CW (germinal and laminated layers) was carefully separated from host tissue with forceps and similarly stored in Trizol. Total RNA was isolated following the manufacturer's instructions. In the case of the CW, the extracted RNA was precipitated with isopropanol in the presence of a high salt solution (0.8 M sodium citrate and 1.2 M NaCl), as recommended for materials containing high levels of polysaccharide. This modification avoids contamination of the RNA with co-precipitating carbohydrates.

2.2. Conventional cDNA library preparation

Total PS RNA (100 μg from a single highly fertile cyst) was used to prepare a cDNA library in the pSPORT vector (GibcoBRL) following the manufacturer's instructions. Briefly, reverse transcription with SuperScript II was primed with a *Not*I-restriction site-tagged oligo-$(dT)_{15}$. After second strand synthesis with *E. coli* DNA polymerase I in combination with *E. coli* RNase H and *E. coli* DNA ligase, cDNA was blunt-ended with T4 DNA polymerase. It was then ligated to *Sal*I adapters, digested with *Not*I and size-fractionated by column chromatography. The fractions containing

cDNAs ≥ 500 bp were subsequently ligated to *Not*I–*Sal*I-cut pSPORT1 and electroporated into *E. coli* (ElectroMAX DH10B cells, GibcoBRL).

2.3. Oligo-capped and spliced leader cDNA libraries

The GeneRacer kit (Invitrogen) was used as shown in Fig. 1 to ligate an RNA oligonucleotide to the 5′ end of the originally capped RNA population from PS, pepsin-treated PS and CW. In all cases, 1 μg of total RNA was treated with calf intestinal phosphatase to dephosphorylate all non-protected polynucleotides (i.e. genomic DNA, ribosomal and mitochondrial RNAs and truncated mRNAs). Then, tobacco acid pyrophosphatase was used to cleave the mRNA cap, exposing a reactive phosphate group at the 5′ end of cap-bearing mRNAs [17]. This allowed them to ligate to the GeneRacer linker oligonucleotide (5′-CGACUGGAGCACGAGGACACUGACAUGGACUGAAGGAGUAGAAA-3′). The ligated RNA was reverse transcribed with the tagged oligo-dT described above and SuperScript II. For amplification of oligo-capped cDNAs, a primer specific for the tag (*Not*I primer: 5′-TAGATCGCGAGCGGCCGCCCTTT-3′) was used together with a primer against the ligated linker (GeneRacer 5′ nested primer: 5′-GGACACTGACATGGACTGAAGGAGTA-3′). PCR was carried out for 15 cycles (1 min at 94 °C, 1 min at 68 °C, 5 min at 72 °C) using *Taq* DNA polymerase (Qiagen). After removal of A-overhangs with *Pfu* DNA polymerase (Stratagene), the products were directionally cloned to *Not* I–*Sal* I-cut pSPORT1 following the protocol outlined in Section 2.2.

Trans-spliced transcripts were also amplified from total RNA (subjected to dephosphorylation and GeneRacer linkage), using the *Not*I primer and a primer specific for *Echinococcus* SL (5′-CACCGTTAATCGGTCCTTACCTT-3′), as described by Brehm et al. [13] employing 20 PCR cycles (1 min at 94 °C, 1 min at 57 °C, 5 min at 72 °C). SL-amplified cDNAs were similarly cloned in pSPORT1.

2.4. Library sequencing and analysis

The libraries were plated out and random colonies picked for expressed sequence tag (EST) sequencing.

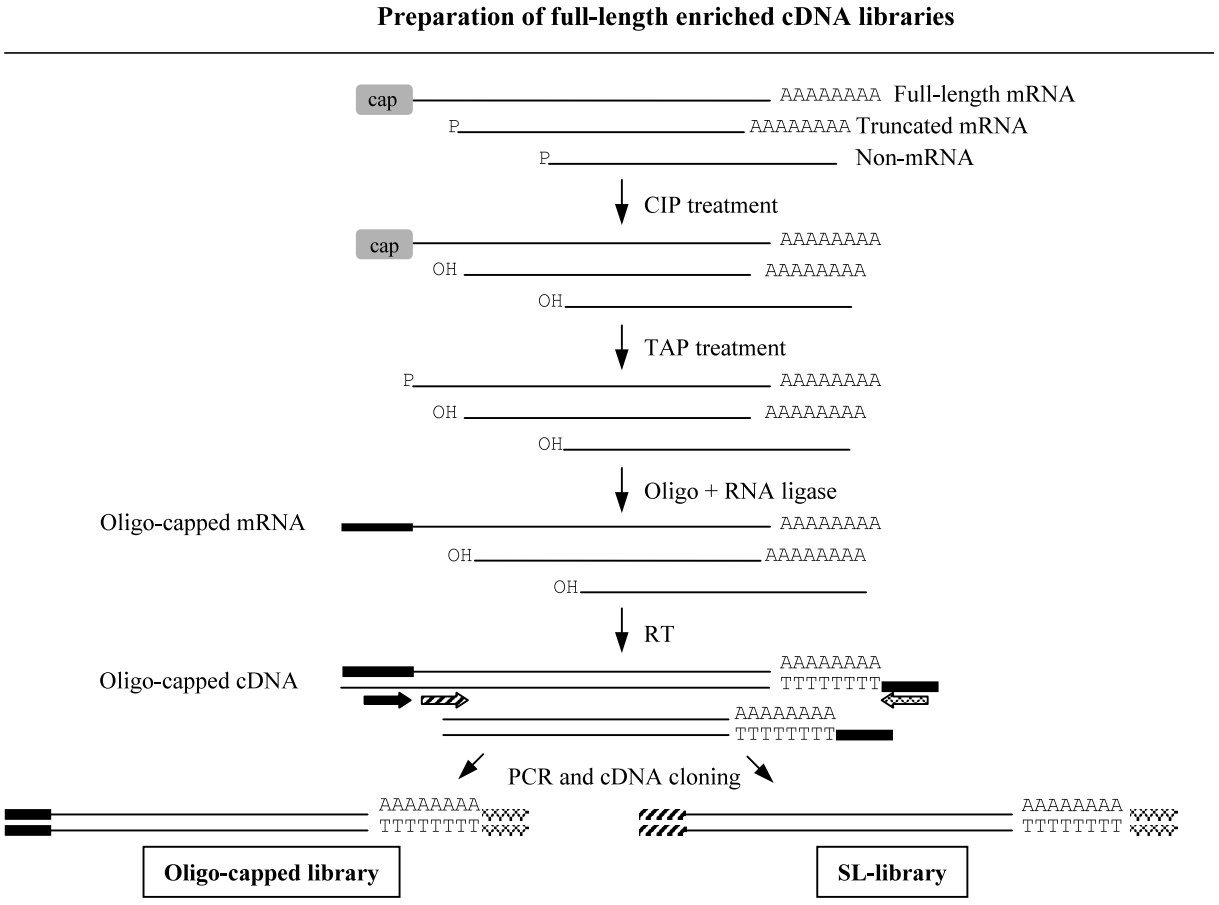

Fig. 1. Schematic representation of the construction of oligo-capped and spliced leader (SL) cDNA libraries. CIP, calf intestinal alkaline phosphatase; TAP, tobacco acid pyrophosphatase; RT, reverse transcriptase. Block arrows represent PCR primers against: ➡ the oligo-cap; ➡ the SL sequence; ➡ the oligo-dT tag. See Section 2 for details of the procedures.

Briefly, isolated colonies were grown overnight in 96-well plates. Cloned cDNAs were amplified from these cultures using vector primers (M13 forward and reverse). After treatment with shrimp alkaline phosphatase and exonuclease I, the PCR products were directly used for single-pass 5′-sequencing of the cloned cDNAs (T7 primer, BigDye terminator cycle sequencing ready reaction, Applied Biosystems).

The sequences were deposited in the National Centre for Biotechnology Information (NCBI) database of expressed sequence tags (dbEST) as ESTs (GenBank accession numbers BI243990–BI244549 and BF642954–BF643192). They were also clustered with the *E. granulosus* sequences already present in GenBank; the resulting clusters were grouped in EchinoBASE and can be retrieved from: http://nema.cap.ed.ac.uk/seq_tables/echinococcus/echinobase.php. This database was prepared with the tools developed at the University of Edinburgh to set up NEMBASE [18], and includes the results of BLAST analyses of the consensus sequence from each cluster against nonredundant GenBank nr and dbEST databases [19]. The sequences from individual ESTs were searched for open reading frames using MACVECTOR 7.0 software (Oxford Molecular); and for the presence of putative signal sequences with the SignalP algorithm [20], obtained from the EXPASY Molecular Biology Server (http://www.expasy.ch/).

3. Results and discussion

3.1. Mitochondrial RNA transcripts in conventional cDNA library

A PS cDNA library of 3×10^5 independent clones, 67% of which contained inserts ≥ 500 bp, was prepared by oligo-dT-primed first-strand synthesis. Numerous *E. granulosus*-specific cDNAs were identified by random EST sequencing of about 250 clones; but, surprisingly, a single transcript predominated (more than 60% of clones) which proved to be the large subunit (LSU) of mitochondrial rRNA (Fig. 2A; Table 1). In all clones for which sequence extended through the 3′ end (89/146) there was a poly-A tract, generally longer than the oligo-$(dT)_{15}$ used in cDNA synthesis (65/89 contained 18–30 A residues). Mapping onto the known mitochondrial genome of *E. granulosus* revealed that the poly-A tract was positioned precisely at the 3′ end of the LSU product, and there was no A-rich genomic stretch in this region which could have caused mis-priming (Fig. 2B). Furthermore, a conserved heptamer motif for transcription termination in animal mitochondria (the 'rRNA termination box', [21]) was identified 21 nt downstream of the LSU gene, inside the contiguous $tRNA^{Cys}$ gene. Thus, the extremely abundant LSU rRNA transcript appears to be polyadenylated in *E. granulosus*. Similar polyadenylation of mitochondrial LSU rRNA has been reported for other organisms, notably *Drosophila* [22] and the platyhelminth *Fasciola hepatica* [23].

The conventional library suffered from further deficiencies: up to 8% of inserts were other polyadenylated mitochondrial transcripts, both small subunit (SSU) rRNA and various mRNAs, and very few of the non-mitochondrial transcripts showed clear 5′ initiation codons. Finally, a small proportion of clones derived from bovine contamination, mainly from genomic DNA as revealed by database searches (Table 1).

Our results show that mitochondrial transcripts predominate among PS cDNAs; and that such transcripts—either rRNAs or mRNAs—are polyadenylated. Where high levels of LSU rRNA have been noted, its presence has generally been interpreted in a developmental context. For example, in *Drosophila* and *Xenopus* embryos mitochondrial rRNAs are both associated with the germ plasma and localised outside mitochondria [24–27]. These studies suggested that polyadenylation mediates the transport of rRNAs across the organelle membrane. Interestingly, in embryos of Planaria, a free-living flatworm, rRNAs are localised in blastomeres that contribute to the formation of the larvae, and not to germline development [28]. This observation in a member of the same phylum as *Echinococcus* suggests that the mitochondrial rRNAs in the latter species could derive from the developing PS, which result from the proliferation of undifferentiated cells at the germinal layer. Further studies would be necessary to confirm the striking predominance of the LSU over the SSU (25-fold excess), as well as to analyse the significance of such high levels of these transcripts for the biology of *E. granulosus* PS.

3.2. Analysis of oligo-capped and spliced-leader libraries

The presence of a major polyadenylated mitochondrial RNA transcript dictated the use of a new approach for cDNA library construction. To restrict our libraries to nuclear-encoded mRNA species, we adapted a technique for the Rapid Amplification of cDNA Ends (RNA-ligase mediated RACE) that excludes all polynucleotides lacking the 7-methyl-guanosine (7MG) cap at the 5′ end of full-length mRNAs. A specific oligo is ligated to the 5′ end, allowing PCR amplification in combination with an oligo-dT primer [17]. The rationale behind the choice of this strategy was that the 7MG cap is exclusively present in RNA polymerase II products; and, thus, would be absent from mitochondrial RNA polymerase-derived transcripts. In addition, the method excludes contamination from ribosomal RNA and genomic DNA (host- or parasite-derived); and favours full-length cDNAs.

In parallel, we made use of the newly-described *Echinococcus* SL, a 36-nt sequence present on an

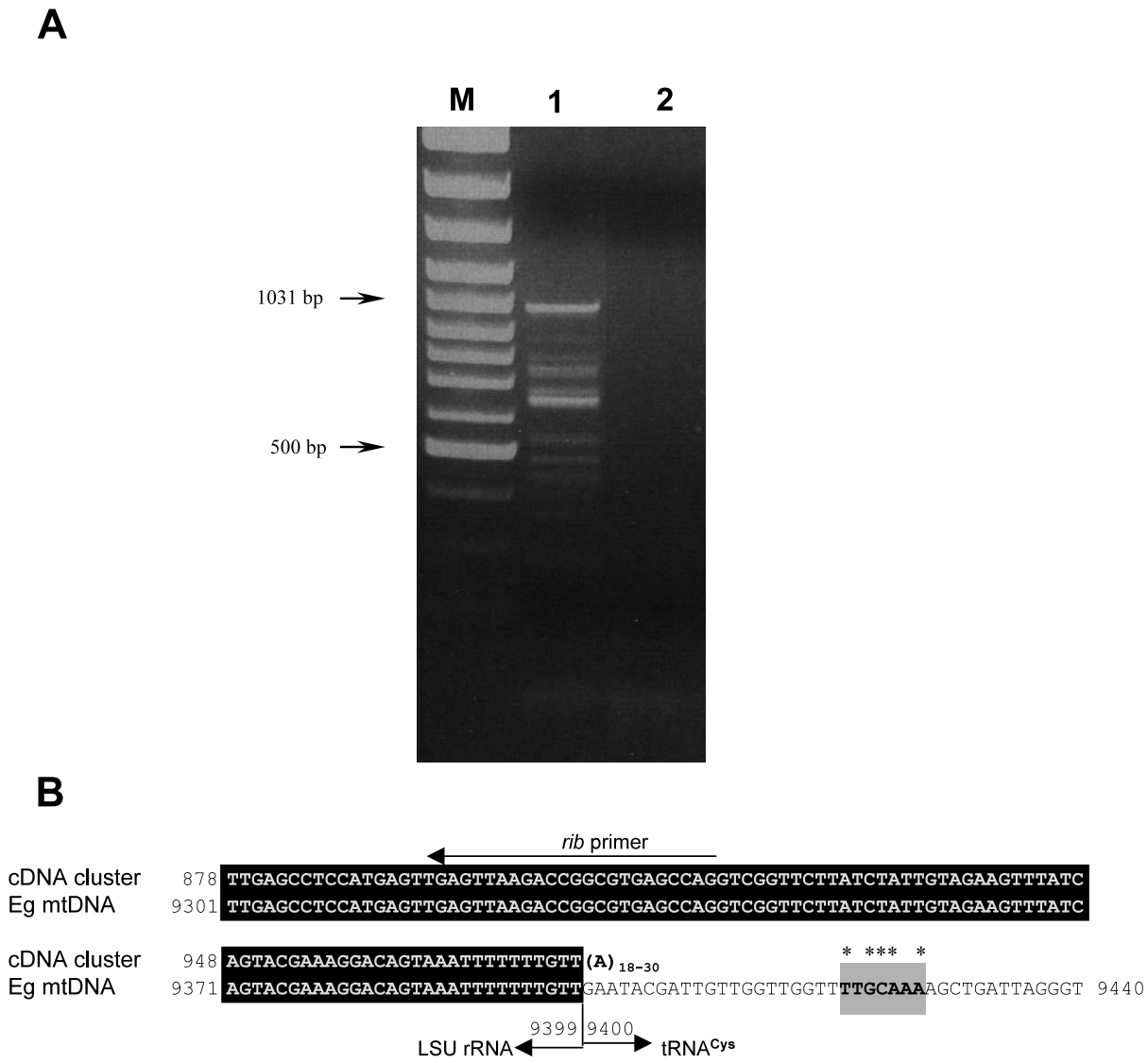

Fig. 2. A: Agarose gel electrophoresis of PCR amplified library of conventional oligo-dT primed cDNA showing dominance of LSU mitochondrial rRNA transcripts. An aliquot of the library (50–500 clones) was amplified with a forward vector primer and a reverse primer specific for the LSU rRNA gene (*rib*: 5′-CTGGCTCACGCCGGTCTTAACTCA-3′; see Fig. 2 B). PCR yielded a range of products of up to about 1000 bp, which is the size expected for a cDNA corresponding to the full-length LSU rRNA (lane 1). No product was obtained with a similar aliquot from the oligo-capped library (lane 2). M: 100 bp DNA ladder (MBI, Fermentas). B: Alignment of cDNA and genomic mitochondrial sequences towards the 3′ end of the LSU rRNA gene, showing the proposed region of polyadenylation. The cDNA sequence represents a cluster of 89 clones for which sequence read through to the poly-A tail. The box indicates the position of the putative rRNA transcription termination box [21] located within the contiguous tRNACys gene; and the asterisks the conserved residues from the consensus heptamer (TGGCAGA). Numbering on the gene sequence refers to the mitochondrial genome of *E. granulosus* (GenBank accession no AF297617; the LSU rRNA gene is located between positions 8423 and 9399; the $tRNA^{Cys}$ gene between 9400 and 9462). Numbering on the cDNA sequence refers to the cluster corresponding to the LSU rRNA in EchinoBASE (EGC00077).

estimated 25% of transcripts [13]. *Trans*-splicing transfers the 5′ portion of a small nuclear RNA species, which includes a distinct cap structure, 2,2,7-trimethylguanosine (TMG) [29]. As originally described in *Caenorhabditis elegans* [30,31], the TMG cap deriving from the donor molecule appears to be retained on *trans*-spliced echinococcal transcripts [13].

We prepared oligo-capped and SL-libraries using RNA not only from PS but also from pepsin-treated PS and from CW. Pepsin increases PS metabolic activity, being a signal naturally encountered by PS upon ingestion in the dog. Exposure to pepsin may up-regulate parasite genes that are important in interaction with the host. The CW, in turn, is the part of the larva directly exposed to the host; it had not been previously included due to the difficulty of preventing host contamination.

Each library was prepared from 1 μg of total RNA, and contained 10^5–10^6 independent clones. Some 80% of the cDNAs in the PS libraries, and 97% in the CW

Table 1
Comparative analysis of cDNA libraries

Library	Number of ESTs	Mit rRNAs	Mit mRNAs	Host (gDNA + cDNA)	Insert size[a] (Median, range) bp	Putative Start Met + ORF[b]	Possible Start Met from SL[c]	Putative Signal Sequence[d]
Conv-dT	248	LSU: 150 (61%) SSU: 6 (2%)	15 (6%)	10 (4%)	700 (350–3000)	5 (7%)[e]	n/a	0
PSU-GR	95	0	0	0	850 (350–1800)	81 (85%)	n/a	12 (13%)
PSU-SL	95	0	0	0	700 (350–1800)	87 (92%)	47 (54%)	26 (27%)
PSP-GR	90	0	0	0	750 (350–1800)	76 (84%)	n/a	14 (16%)
PSP-SL	86	0	0	0	750 (350–1800)	66 (77%)	30 (45%)	10 (12%)
CW-GR	89	0	0	4 (5%)	950 (400–2800)	54 (61%)	n/a	10 (11%)
CW-SL	104	0	0	0	1050 (400–2800)	92 (88%)	42 (46%)	17 (16%)

Abbreviations: gDNA, genomic DNA; n/a, not applicable. Library codes: PSU, Untreated protoscoleces; GR, GeneRacer (oligo-capped); PSP, Pepsin-treated protoscoleces; SL, Spliced leader; CW, Cyst Wall; Conv-dT, Conventional oligo-dT primed.

[a] The size of the cDNA inserts was estimated by agarose gel electrophoresis analysis of the corresponding PCR products.

[b] An ATG codon towards the 5′ end of an EST was considered to be a putative starting Met if it was followed by an open reading frame of at least 99 nt. When available, similarity with an orthologous protein was also considered.

[c] Putative starting codons contributed by the echinococcal SL. Percentage is of sequences with identified start sites.

[d] A signal sequence was considered to be present when at least two parameters scored positive using the SignalP algorithm [19].

[e] Percentage is of sequences other than mitochondrial and host ESTs.

libraries, were found to be bigger than 500 bp. Approximately 100 randomly selected clones from each library were PCR-amplified and single-pass sequenced from the 5′ end to assess the quality of the libraries. Overall, a low level of redundancy was found, in spite of having used a PCR approach for cDNA synthesis, with >65% of sequenced clones corresponding to different cDNAs. The global features of sequence analysis are presented in Table 1.

Contamination with cDNAs derived from mitochondrial and ribosomal transcripts was virtually excluded (0/559 sequenced clones). Some bovine-derived clones were found in the CW oligo-capped library (4/89), but these were all cDNAs showing that host genomic contamination was avoided. A level of 5% host cDNA is not surprising in view of the fact that the CW forms an intimate biological interface between host and parasite [4].

3.3. Enrichment of full-length cDNAs

The primary criterion we used to judge the representation of full-length cDNAs in each library was the presence of a putative initiation codon, arbitrarily assigned where an ATG was followed by at least 99 nt of open reading frame. Only ~7% of the parasite cDNA clones from the conventional library fulfilled this condition, but in the oligo-capped and SL-libraries, the majority (61–92%) of cDNAs met this criterion. In ~60% of instances, similarity with a protein sequence in the database allowed comparison of the putative start methionines and to test the validity of our approach. In fact, inconsistencies were noted in a small number of cases (<1.5%), and only for cDNAs showing significant but not particularly high similarities ($10^{-5} < P < 10^{-15}$).

In all flatworms analysed so far (*Echinococcus*, *Schistosoma* and *Fasciola* spp.), the SL contains an AUG trinucleotide which offers a potential translation initiation point. We, therefore, identified the putative open reading frames in the SL-bearing cDNAs, which are or are not in frame with the SL AUG possible start site. Previous studies, on highly conserved protein sequences from *E. multilocularis* and *S. mansoni* found instances in which the SL codon was used as the translation start [13,32]. We confirmed that some *E. granulosus* mRNAs are also likely to use the SL AUG, such as those corresponding to a peptidyl-prolyl isomerase E (GenBank accession no BI244411 and BI244429) and a lysosomal ATPase subunit (GenBank accession no BI244083). In these cases, translation from downstream AUGs would result in products lacking a considerable part of the conserved sequence. Using our broader criterion (open reading frame ≥99 nt following the start codon), ~43% of the SL-containing messages (73/168) could start at the SL AUG. This figure may well be an overestimate, due to the high proportion of such transcripts bearing no homology to other sequences in the databases. However, our data are consistent with *E. multilocularis* in which the SL-AUG was out of frame of the actual start in ~50% of the transcripts analysed (6/11) [13].

Identification of the first in-frame AUG allowed us to analyse the length of 5′ untranslated regions (UTRs) in *E. granulosus* mRNAs. The majority of 7MG-capped mRNA clones, represented in the oligo-capped library, showed intervals of 21–60 nt between the start of the mRNA and the suggested initiator codon (106/170, 62%) as shown in Fig. 3A. Some 33 clones (19%) had ≤20 nt, and only four clones (2%) had >200 nt of 5′ UTR. The length distribution of 5′ UTR in SL-bearing mRNAs was very different. Not only did 43% have no 5′ UTR at all, initiating from the SL sequence, but in a further 39% (65/168) there were ≤20 nt separating the 3′ end of the SL from the AUG codon (Fig. 3B).

Where the 5′ end of clones showed homology to known *E. granulosus* cDNAs, it was possible to compare data from the new libraries with existing information. For example, from the oligo-capped libraries we have isolated putative full-length transcripts of *E. granulosus* actin [33], thioredoxin peroxidase [34], cyclophilin [35], elongation factor-1∂ [36], glutathione S-transferase [37], P-29 [38], myophilin [39] and heat shock protein 70 [GenBank accession no U26448]. All but two of the existing sequences (P-29 and myophilin) are truncated, with new clones providing 43–310 nt additional 5′ sequence. The 5′ end of P-29, previously determined by 5′- RACE, is identical to the EST sequence. Only in the case of myophilin is the transcript from the oligo-capped library shorter, by 43 nt. The existing full-length cDNA was isolated from an adult library; thus, it is possible that the transcription of this gene is initiated at different sites in the larval and adult stages.

In addition, the length of cDNA inserts was assessed. As shown in Table 1, the libraries were not significantly different in terms of median size. Because PCR amplification was used for both oligo-capped and SL-libraries, it seems likely that large transcripts are under-represented, as has previously been noted [14,16]. It is possible that the size distribution of such libraries can be improved by modifying the conditions under which the PCR reaction is performed [40]. However, we observed that the inserts in both CW libraries are consistently longer, indicating that the insert size is also determined by the quality of the starting material.

3.4. Signal-sequence bearing cDNAs

A further measure of full-length cDNA library quality is the proportion of inserts encoding potential signal sequences. A signal sequence was considered present

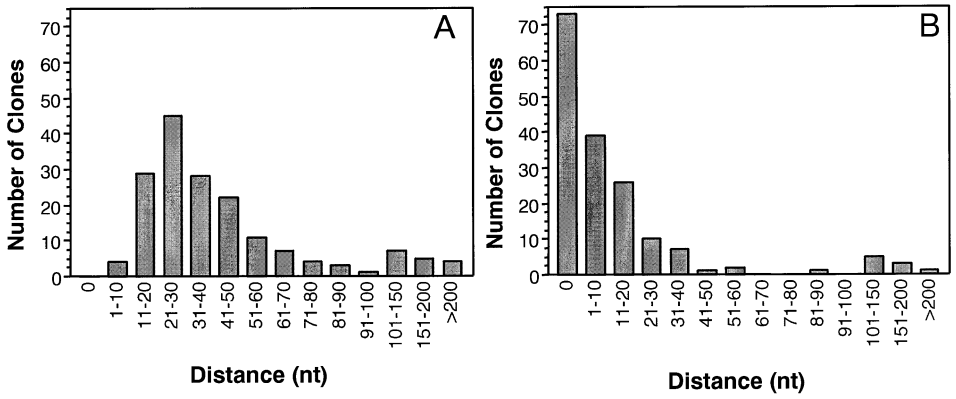

Fig. 3. Length of 5′ untranslated regions (UTRs) in full-length cDNA clones derived from (A) 170 oligo-capped clones and (B) 168 spliced-leader mRNA clones. Cluster analysis was used to ensure only a single clone from any one gene was analysed. In 11/159 oligo-capped clusters, two alternative transcript lengths were observed, and both were scored. Distances are given as number of nucleotides from the ligated oligo or the 3′ end of the SL to the putative AUG start codon. In (B), zero represents the clusters in which the terminal trinucleotide AUG of the SL is capable of acting as the start site for translation. Start sites were attributed if they initiated open reading frames of at least 99 nucleotides as described in the text.

when at least two parameters scored positive using the SignalP algorithm [20]. Not a single cDNA from the conventional library encoded an N-terminal sequence predicted to act as a signal sequence. In contrast, a significant proportion of cDNA sequences from the oligo-capped and SL sources was found to contain putative signal peptides (11–27% in the different libraries), corresponding to 60 different proteins. These clones are particularly interesting because secreted and membrane proteins are likely to represent the most critical genes involved in host–parasite cross-talk. A more detailed analysis of such genes has now commenced (Fernández C., Parkinson J. & Maizels R.M., work in progress).

3.5. Non-overlapping sets of cDNAs from oligo-capped and trans-spliced libraries

Finally, we compared the repertoire of cDNAs recovered from each library. Surprisingly, only one cDNA of the 274 derived from oligo-capped libraries was found to contain the SL sequence, despite the estimate that 25% of all mRNAs are *trans*-spliced [13]. This indicates that echinococcal TMG-capped mRNAs are processed most inefficiently (>50-fold less) by the enzymes selected in the oligo-capping protocol. It is instructive to compare this finding with data from *C. elegans*, in which 65% of cDNAs isolated from several oligo-capped libraries were found to contain SL-sequences (Thierry-Mieg et al. 13th International *C. elegans* meeting, 2001). This figure correlates well with the estimated proportion of mRNAs whose maturation involves *trans*-splicing in this organism [41]. In addition, work currently in progress in our laboratory has similarly identified a significant proportion (50%) of SL-transcripts in an oligo-capped library from the parasitic nematode *Nippostrongylus brasiliensis* (Y. Harcus, C. Fernández and R.M. Maizels, unpublished).

These data indicate that the results in *E. granulosus* cannot be attributed to a difference in the relative susceptibilities of 7MG and TMG caps to the pyrophosphatase treatment. Furthermore, they suggest that the *trans*-spliced echinococcal transcripts may bear structural features that are absent from the nematode SL-mRNAs. In this context, it is interesting to mention that the RNA ligase-mediated RACE strategy does not appear to amplify the 5′ ends of *T. brucei* cDNAs (Jeremy Mottram, personal communication). The SL-RNA cap in trypanosomatids is unique among eukaryotes and consists of 7MG followed by four methylated nucleotides (cap 4) [42]. Since this cap is susceptible to tobacco acid pyrophosphatase hydrolysis [43], the lack of success of the RACE strategy is likely due to the 5′ hypermethylated mRNAs not being suitable substrates for T4 RNA ligase. Further characterisation of the *E. granulosus* SL RNA is necessary to elucidate whether such structural features are indeed interfering with oligo-capping of *trans*-spliced mRNAs. Finally, the echinococcal SL is homologous to the one of trematodes and differs substantially from the nematode SL [13]; thus, it is tempting to speculate that the lack of overlap of our oligo-capped and *trans*-spliced libraries may reflect an unusual and distinctive feature of flatworm SL RNAs.

In conclusion, our comparison of oligo-capped and SL-libraries shows that both approaches provide similar enhancement of cDNA library quality, with high proportions of full-length transcripts including signal sequence-encoding genes, minimal levels of contamination with mitochondrial and ribosomal RNA or genomic material, and low degrees of redundancy. Some limitations have been noted, particularly in representation of larger mRNA species. Moreover, the two

methods amplify non-overlapping populations of mRNAs in *E. granulosus* probably due to distinct properties of its SL RNA; thus, it is likely that both techniques may also need to be combined to obtain a broader picture of the genes expressed by other parasitic flatworms.

Acknowledgements

CF is supported by a Wellcome Trust International Travelling Fellowship; WFG by an MRC Postdoctoral Fellowship; PL by a Wellcome Trust Prize Fellowship; and RMM's laboratory by a Wellcome Trust Programme Grant. We thank John Parkinson for invaluable assistance with sequence analysis; and Gustavo Salinas and Alvaro Díaz for fruitful and encouraging discussions throughout the progress of this work.

References

[1] Walder J.A., Eder P.S., Engman D.M., et al. The 35-nucleotide spliced leader sequence is common to all trypanosome messenger RNAs. Science 1986;233:569–71.
[2] Benne R., Van den Burg J., Brakenhoff J.P.J., Sloof P., Van Boom J.H., Tromp M.C.. Major transcript of the frameshifted *coxII* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell 1986;46:819–26.
[3] Muller R.. Worms and Human Disease, 2nd edn. Wallingford: CABI Publishing, 2002.
[4] Thompson R.C.A.. *Echinococcus* and Hydatid Disease, 2nd edn. Wallingford: CABI Publishing, 1995.
[5] Okayama H., Berg P.. High-efficiency cloning of full-length cDNA. Mol Cell Biol 1982;2:161–70.
[6] Gems D.H., Ferguson C.J., Robertson B.D., Page A.P., Blaxter M.L., Maizels R.M.. An abundant, *trans*-spliced mRNA from *Toxocara canis* infective larvae encodes a 26 kDa protein with homology to phosphatidylethanolamine binding proteins. J Biol Chem 1995;270:18517–22.
[7] Yenbutr P., Scott A.L.. Molecular cloning of a serine proteinase inhibitor from *Brugia malayi*. Infect Immun 1995;63:1745–53.
[8] Gregory W.F., Blaxter M.L., Maizels R.M.. Differentially expressed, abundant *trans*-spliced cDNAs from larval *Brugia malayi*. Mol Biochem Parasitol 1997;87:85–95.
[9] Donelson J.E., Zeng W.. A comparison of *trans*-RNA splicing in trypanosomes and nematodes. Parasitol Today 1990;6:327–34.
[10] Nilsen T.W.. *Trans*-splicing of nematode premessenger RNA. Annu Rev Microbiol 1993;47:413–40.
[11] Blaxter M., Liu L.. Nematode spliced leaders—ubiquity, evolution and utility. Int J Parasitol 1996;26:1025–33.
[12] Davis R.E.. Spliced leader RNA *trans*-splicing in metazoa. Parasitol Today 1996;12:33–40.
[13] Brehm K., Jensen K., Frosch M.. mRNA trans-splicing in the human parasitic cestode *Echinococcus multilocularis*. J Biol Chem 2000;275:38311–8.
[14] Suzuki Y., Yoshitomo-Nakagawa K., Maruyama K., Suyama A., Sugano S.. Construction and characterization of a full length-enriched and a 5′- end-enriched cDNA library. Gene 1997;200:149–56.
[15] Yudate H.T., Suwa M., Irie R., et al. HUNT: launch of a full-length cDNA database from the Helix Research Institute. Nucl Acids Res 2001;29:185–8.
[16] Watanabe J., Sasaki M., Suzuki Y., Sugano S.. FULL-malaria: a database for a full-length enriched cDNA library from human malaria parasite, *Plasmodium falciparum*. Nucl Acids Res 2001;29:70–1.
[17] Maruyama K., Sugano S.. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene 1994;138:171–4.
[18] Parkinson J., Whitton C., Guiliano D., Daub J., Blaxter M.. 200000 nematode expressed sequence tags on the net. Trends Parasitol 2001;17:394–7.
[19] Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J.. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[20] Nielsen H., Engelbrecht J., Brunak S., von Heijne G.. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997;10:1–6.
[21] Valverde J.R., Marco R., Garesse R.. A conserved heptamer motif for ribosomal RNA transcription termination in animal mitochondria. Proc Natl Acad Sci USA 1994;91:5368–71.
[22] Benkel B.F., Duschesnay P., Boer P.H., Genest Y., Hickey D.A.. Mitochondrial large ribosomal RNA: an abundant polyadenylated sequence in *Drosophila*. Nucl Acids Res 1988;16:9880.
[23] Zurita M., Bieber D., Ringold G., Mansour T.E.. cDNA cloning and gene characterization of the mitochondrial large subunit (LSU) rRNA from the liver fluke *Fasciola hepatica*. Evidence of heterogeneity in the fluke mitochondrial genome. Nucl Acids Res 1988;16:7001–12.
[24] Ding D., Whittaker K.L., Lipshitz H.D.. Mitochondrially encoded 16S large ribosomal RNA is concentrated in the posterior polar plasm of early *Drosophila* embryos but is not required for pole cell formation. Dev Biol 1994;163:503–15.
[25] Kobayashi S., Amikura R., Mukai M.. Localization of mitochondrial large ribosomal RNA in germ plasm of *Xenopus* embryos. Curr Biol 1998;8:1117–20.
[26] Amikura R., Kashikawa M., Nakamura A., Kobayashi S.. Presence of mitochondria-type ribosomes outside mitochondria in germ plasm of *Drosophila* embryos. Proc Natl Acad Sci USA 2001;98:9133–8.
[27] Kloc M., Bilinski S., Chan A.P., Etkin L.D.. Mitochondrial ribosomal RNA in the germinal granules in *Xenopus* embryos revisited. Differentiation 2001;67:80–3.
[28] Sato K., Sugita T., Kobayashi K., et al. Localization of mitochondrial ribosomal RNA on the chromatoid bodies of marine planarian polyclad embryos. Dev Growth Differ 2001;43:107–14.
[29] Blumenthal T., Thomas J.. *Cis* and *trans* mRNA splicing in *C. elegans*. Trend Genet 1988;4:305–8.
[30] Liou R.-F., Blumenthal T.. *trans*-spliced *Caenorhabditis elegans* mRNAs retain trimethylguanosine caps. Mol Cell Biol 1990;10:1764–8.
[31] Van Doren K., Hirsh D.. mRNAs that mature through *trans*-splicing in *Caenorhabditis elegans* have a trimethylguanosine cap at their 5′ termini. Mol Cell Biol 1990;10:1769–72.
[32] Davis R.E., Hardwick C., Tavernier P., Hodgson S., Singh H.. RNA *trans*-splicing in flatworms. Analysis of *trans*-spliced mRNAs and genes in the human parasite, *Schistosoma mansoni*. J Biol Chem 1995;270:21813–9.
[33] da Silva C.M., Ferreira H.B., Picon M., et al. Molecular cloning and characterization of actin genes from *Echinococcus granulosus*. Mol Biochem Parasitol 1993;60:209–19.
[34] Salinas G., Fernández V., Fernández C., Selkirk M.E.. *Echinococcus granulosus*: cloning of a thioredoxin peroxidase. Exp Parasitol 1998;90:298–301.
[35] Lightowlers M.W., Haralambous A., Rickard M.D.. Amino acid sequence homology between cyclophilin and a cDNA-cloned

antigen of *Echinococcus granulosus*. Mol Biochem Parasitol 1989;36:287–9.

[36] Margutti P., Ortona E., Vaccari S., et al. Cloning and expression of a cDNA encoding an elongation factor 1beta/delta protein from *Echinococcus granulosus* with immunogenic activity. Parasite Immunol 1999;21:485–92.

[37] Fernández V., Chalar C., Martínez C., Musto H., Zaha A., Fernández C.. *Echinococcus granulosus*: molecular cloning and phylogenetic analysis of an inducible glutathione S-transferase. Exp Parasitol 2000;96:190–4.

[38] González G., Spinelli P., Lorenzo C., et al. Molecular characterization of P-29, a metacestode-specific component of *Echinococcus granulosus* which is immunologically related to, but distinct from, antigen 5. Mol Biochem Parasitol 2000;105:177–84.

[39] Martin R.M., Chilton N.B., Lightowlers M.W., Gasser R.B.. *Echinococcus granulosus* myophilin—relationship with protein homologues containing 'calponin-motifs'. Int J Parasitol 1997;27:1561–7.

[40] Piao Y., Ko N.T., Lim M.K., Ko M.S.. Construction of long-transcript enriched cDNA libraries from submicrogram amounts of total RNAs by a universal PCR amplification method. Genome Res 2001;11:1553–8.

[41] Zorio D.A.R., Cheng N.N., Blumenthal T., Spieth J.. Operons as a common form of chromosomal organization in *C. elegans*. Nature 1994;372:270–2.

[42] Bangs J.D., Crain P.F., Hashizume T., McCloskey J.A., Boothroyd J.C.. Mass spectrometry of mRNA cap 4 from trypanosomatids reveals two novel nucleosides. J Biol Chem 1992;267:9805–15.

[43] Mair G., Ullu E., Tschudi C.. Cotranscriptional cap 4 formation on the *Trypanosoma brucei* spliced leader RNA. J Biol Chem 2000;275:28994–9.