Orphon spliced-leader sequences form part of a repetitive element in *Angiostrongylus cantonensis*

G. W. P. Joshua*, F. B. Perler¹ and C. C. Wang²

Institute of Biomedical Sciences, Academia Sinica, Taipei 11529, Taiwan, Republic of China, ¹New England Biolabs Inc., 32 Tozer Road, Beverly, MA 01915, USA and ²University of California, San Francisco, CA 94143, USA

Received October 19, 1994; Revised and Accepted February 7, 1995

GenBank accession nos U13190 and U13191

**ABSTRACT**

In nematodes, the 22 nucleotide (nt) spliced leader (SL) is normally encoded by a multi-copy, tandemly reiterated SL gene and is trans-spliced from SL-RNA onto the 5' end of a subset of mRNAs. We have found that the SL is also encoded at multiple (>100) orphon genomic sites in the parasitic nematode *Angiostrongylus cantonensis*. At these sites the sequence forms part of a 198 bp repetitive element (designated con-198). Transcription from two genomic elements that contain the con-198 sequence has been characterised. At one element (G-2) an ~850 nt RNA with an internal SL is transcribed. At the other (G-1), transcription takes place 3 kb downstream of the con-198 sequence.

**INTRODUCTION**

*Angiostrongylus cantonensis* is a parasitic nematode normally found, at the adult stage, as a lung worm in rats. After oral infection it has a migratory period that includes infection of the brain. When it infects humans, development is arrested at the adult stage, as a lung worm in rats. After oral infection.

Like the mRNAs of other nematodes (2-5), kinetoplastid protozoa (6), trematodes (7) and *Euglena* (8), mRNAs of *A. cantonensis* include in their maturation the trans-splicing of a 5' untranslated leader sequence. In nematodes this process occurs on a sub-set of mRNAs (3). Nematode genes may also contain introns, necessitating both cis- and trans-splicing of mRNAs. In all nematodes studied to date the trans-spliced leader sequence (SL) is 22 nucleotides (nt) long and is absolutely conserved (5'-GGTTTAATTACCCAAGTTTGAG-3'). A second SL sequence is thought to occur in *A. cantonensis* as it does in other nematodes.

It possesses an SM antigen binding site and has a secondary structure which resembles that predicted for SL-RNAs of both kinetoplastids and nematodes (14). Thus, trans-splicing of the SL sequence is thought to occur in *A. cantonensis* as it does in other nematodes.

The SL sequence of both kinetoplastids and nematodes has been found at genomic locations other than the tandemly reiterated SL-genes participating in trans-splicing. These have been termed 'orphon' sequences (15). In kinetoplastids, orphon SL sequences are associated with DNA homologous to the intergenic regions of the SL gene repeat. In addition, the SL gene repeat may be interrupted by regularly interspersed transposable inserted elements; the CRE1 element in *Crithidia fasciculata* (16), and the SLACS (17,18) and MAE (19) elements in *Trypanosoma brucei*. In contrast, orphon SL sequences in nematodes are not associated with other regions of the SL gene repeat. In *Brugia malayi*, no transcription of orphon SL sequences is detected (4) whereas in *Onchocerca volvulus* transcription from several different DNAs with an orphon SL sequence is seen (20).

In this work we initially set out to identify trans-spliced mRNAs, by screening an *A. cantonensis* cDNA library with the anti-sense 22 nt SL sequence to identify full length cDNAs with the trans-spliced leader sequence. We report rather the identification of cDNAs with an internal rather than 5' SL sequence, the isolation of corresponding genomic elements in which the SL sequence forms part of a 198 bp repetitive element, and transcription from these genomic elements.

**MATERIALS AND METHODS**

**Parasites**

The *A. cantonensis* life-cycle was maintained through *Achatina fulica* snails and Sprague Dawley rats. Adult *A. cantonensis* were extracted from rat pulmonary blood vessels 40 days post infection.

**Nucleic acids methods**

Nucleic acids were extracted as described previously (13). A cDNA library was constructed from adult *A. cantonensis* poly(A)* RNA in AGT11 following standard procedures (21) using an Amersham cDNA construction kit with random hexamer

* To whom correspondence should be addressed
primers. A genomic library was constructed from DNA restricted partially with Alul, Rsal and HaeIII. Following EcoR1 linker ligation and methylation, DNA was ligated to λGT11. For sequencing, clones were subcloned into pBluescript KS plasmid (Stratagene) and sequenced by the dideoxy method (22).

Southern blots of genomic or cloned DNA were hybridised in 50% formamide, 1% SDS, 6 x SSC and 100 μg/ml salmon-sperm DNA at 42°C overnight. Blots were washed in 10 x SSC/0.5% SDS, 1 x SSC/0.5% SDS; 0.1 x SSC/0.5% SDS for 20 min each sequentially at room temperature, at 37°C and at 65°C (1 x SSC was 150 mM NaCl and 15 mM sodium citrate).

Northern blots were hybridised in 50% formamide, 2 x Denhardt’s solution, 0.1% SDS, 5 x SSC and 100 μg/ml salmon sperm DNA at 42°C overnight and washed as above. Restriction digestions were performed following the manufacturer’s instructions (New England Biolabs). Restriction fragments were gel purified using silica gel (GeneClean, Bio 101).

Probes were labelled with [α-32P]dCTP by random oligonucleotide labelling (Pharmacia). Primer extension analyses were performed using 1 ng T4 polynucleotide kinase labelled oligonucleotides indicated in figure legends and shown in Figure 1. [Primers: -6(1): 5’-GTCGTACAGGAAGCAATATCTTGCTAGAGG-3’; P-4: 5’-TTTTTTGTGTTGACCTAGC-3’; G-1 3630/3610: 5’-TTGAAATGATTGCAGCTGGT-3’]. These were hybridised to 10-20 μg A.cantonensis RNA at 30°C in 40 mM PIPES (pH 6.4), 1 mM EDTA, 0.4 M NaCl and 80% formamide and extended in 50 mM Tris (pH 7.6), 75 mM KCl, 3 mM MgCl2, 1 mM each dNTP and 10 mM DTT with 200 U Moloney Murine Leukaemia Virus reverse transcriptase (M-MLV-RT; BRL). Extension products were fractionated in 7 M urea—6% acrylamide gels and autoradiographed.

For reverse transcriptase polymerase chain reaction (RT-PCR), 10 pmol 3’ oligonucleotide primers indicated in figure legends and shown in Figure 1 were reverse transcribed with 5 μg A.cantonensis RNA in 50 mM KCl, 20 mM Tris (pH 8.3), 25 mM MgCl2, 1 mM each dNTP and 200 U M-MLV reverse transcriptase. The 5’ primers were: G-2 2205/2224, 5’-GCGGTTAATGCCTCTCTCA-3’; G-2 2604/2624, 5’-TTCTTAGTGGCTGATTCAATC-3’; G-2 2654/2674, 5’-CATTACTGCCTGTGGTTATAT-3’; G-2 2705/2725, 5’-GAAGCATTAAATGCAATAT-3’; G-2 2755/2775, 5’-CCGCTGTCCACATCTTGCAT-3’; G-2 2806/2828, 5’-TCGAGATTATTTCTTACACAA-3’.

The 3’ primers were: con 54, 5’-CTTTGTTTGTTCGACGGCAGGA-3’; G-2 3306/3285, 5’-AACGACCTTCTTGGCAGAATTC-3’; G-2 3346/3326, 5’-CAACGACGAGGCTTTCAGCA-3’.

After addition of 5’ primers (10 pmol), reactions were amplified with 2.5 U Taq polymerase (AmpliTag; Perkin-Elmer Cetus) for 25 cycles of 94°C, 1 min; 55°C, 1 min and 72°C, 2 min. Amplification products were separated in 0.8% agarose, blotted onto nylon membrane and probed with cloned 32P-p-con-198.

RESULTS

Identification of orphon SL genomic sequences

Initially nine cDNA clones were identified and purified after screening a λGT11 cDNA library (~5 x 10⁶ p.f.u.) with an anti-sense 22 nt SL oligonucleotide. These ranged from 1.2 to 4.0 kb and shared substantial (>90%) homology on partial or complete sequencing. In all clones the SL 22 nt sequence was not at the 5’ end. cDNA clones were not polyadenylated and did not possess open reading frames (ORFs) coding for polypeptides longer than 90 amino acids. These features, together with the observation that one cDNA clone (si 6) showed 95% similarity to the genomic clone G-2 (see below) raise the possibility that these cDNA clones resulted from DNA contaminating the RNA used in cDNA library construction. One cDNA, designated sl 233, was used to screen a λGT11 genomic library (~5 x 10⁶ p.f.u.) of A.cantonensis DNA. Fifteen clones were identified, purified and sub-cloned in pBluescript.

Analysis of genomic clones with orphon SL sequence(s)

After partial or complete sequencing of four of the 15 positive genomic clones, two clones, G-1 and G-2, were selected for further analysis. G-2, 3.5 kb long, contained two SL sequences. The upstream of these two SL sequences had one base change, the downstream sequence was perfectly conserved. In addition, the region extending 125 bp upstream and 51 bp downstream of the two SL sequences showed 95% identity. This 198 nt sequence (i.e. including the 22 nt SL sequence) was designated con-198 (Fig. 1) and was found in all cDNA (6) and genomic clones (4) that were completely or partially sequenced. In G-2 (and sl 6) the sequences were designated con-198 (upstream) and con-198 (downstream). In addition, the sequence 5’ of the two con-198 sequences in G-2 showed <30% similarity for 140 nt, but thereafter showed 85% similarity in a 900 nt region. Thus, the con-198 sequence apparently formed part of a degenerate 1.2 kb tandem repeat which had 913 bp of unrelated sequence separating the two repeat units. It is worth noting that examination of the G-2 sequence (GenBank accession no. U13190) shows several features diagnostic of a retrotransposable element (shown in Fig. 1). The con-198 sequence would constitute the long terminal repeats (LTRs) terminating in 5’ TG and 3’ CA dinucleotides. Flanking the LTRs are imperfect 4 bp repeats (5’-AACA; 3’-AGGA) resembling a target site duplication, through these flank both sides of both con-198 sequences. Close to the 5’ putative LTR there is a 20 bp region in which 16 nt are complementary to the 3’ end of the tRNA specified found in Bacillus steaurothermus (23). This might serve as the primer binding site (PBS) for first strand synthesis. Similarly, upstream of the 3’ putative LTR is a 5 bp oligopurine stretch which may serve as the primer binding site (PB*) for second strand synthesis (see Fig. 1).

The clone G-1 was similar to G-2 only in the 198 nt con-198 sequence (Fig. 1). This 4.2 kb genomic clone possessed one con-198 region, close to the 5’ end, in which the SL sequence showed two base changes compared to the conserved nematode SL sequence. Apart from the con-198 sequence G-1 showed no similarity to G-2 except for the 30 nt adjacent to the 3’ end of the con-198 sequence. This 30 nt region shared 85% similarity with the equivalent 30 nt downstream of the 5’ con-198 sequence (G-2) (Fig. 1). The regions adjacent to con-198 (both 5’ and 3’) in other sequenced genomic clones showed 89-91% similarity to regions adjacent to con-198 (upstream) in G-2. The cDNA clones showed similar homology to con-198 (downstream) in G-2.

Southern blot analysis of A.cantonensis DNA digested with a variety of restriction endonucleases and probed with con-198 [cloned following reverse-transcriptase polymerase chain reac-
Figure 1. Partial nucleotide sequences of G-2, showing the upstream and downstream con-198 sequences and intervening sequence; the cloned SL gene [pAcSL, ref (13)]; and G-1 (complete sequences are lodged in GenBank with accession nos U13191 for G-1 and U13190 for G-2). Identity between clones is shown by vertical lines; ends of clones are shown by H. Note that nucleotide numbering, above the line in G-2 and below the line in G-1 starts from the beginning of the con-198 sequence.

In both G-1 and G-2, the con-198 sequence is boxed and the SL sequence is hatched. In the pAcSL sequence, the 3' end of the gene is shown by an oblique arrow. Primer sites for primer extension and PCR (see Materials and Methods and Fig. 5) are marked and arrowed to give orientation. The transcriptional start sites in G-2 and G-1 are bracketed at nucleotides 1126 and 3527 respectively and upstream CAT and TATA boxes are underlined (broken). The retrotransposon features in G-2 are indicated: bold underline shows 4 bp imperfect repeat; and PBS and PBS+ show putative primer binding sites.

We estimated the copy numbers of the con-198 sequence, of the intervening sequence between con-198 (upstream) and con-198 (downstream) (data not shown) by colony hybridisation to the con-198 specific primers—data not shown—showed hybridisation primarily to single bands (Fig. 2, panel 3). Limited hybridisation to other bands was visible and hybridisation to multiple bands occurred when DNA was digested with enzymes that cut more frequently (lanes 3, 5 and 6). Probing with G-2 resulted in hybridisation to multiple bands, and probing with G-1 in hybridisation to one or two bands (Fig. 2, panels 1 and 2). Comparison of panels 1, 2 and 3 in Figure 2 shows that the major hybridisation bands detected by each of the three probes were different. These differences in hybridisation indicate that the non-con-198 regions of G-1 and G-2 occupy different genomic locations. However, DNA digested with each enzyme showed at least one band common to all three probes, which presumably reflects hybridisation to the con-198 sequence. Due to the intensity of the major bands, common bands are faint in some lanes. All three probes gave a radically different pattern of hybridisation compared to the cloned SL gene [pAcSL, ref (13)] from which the trans-spliced SL-RNA is transcribed (Fig. 2, panel 4). Note that hybridisation and washing conditions were too stringent to allow hybridisation of the 22 nt SL sequence only.
Transcription of genomic elements with orphon SL sequences

To confirm transcription from these genomic elements that contain the con-198 sequence, a series of Northern blot experiments was performed. Probes prepared from G-1, G-2 and the cloned con-198 sequence (see Fig. 4A) were hybridised to A. cantonensis total RNA. When the RNA was separated in 1% formaldehyde-agarose, G-2, G-2H (spanning con-198) and the con-198 sequence were dot blotted onto nylon membranes (Schleicher and Schuell) and probed with G-1 (panel 1), G-2 (panel 2), 198 nt con-198 sequence (panel 3) and with the cloned SL gene pAcSL (see 13) (panel 4). Hybridisations were performed as in Materials and Methods, final wash was at 65° C in 0.1 x SSC and 0.5% SDS. Molecular weight markers are shown to the left.

(downstream) in G-2, and of the transcribed region downstream of con-198 in G-1 (see below). These sequences were isolated from the corresponding clones by restriction digestion with the appropriate enzyme (see Fig. 3) and gel-purified. The 32P-labelled probes were hybridised to known amounts of genomic DNA or plasmids containing G-1, G-2 or the con-198 sequence (Fig. 3). By assuming a genomic complexity of 8 x 10^7 bp (the complexity of C. elegans) for A. cantonensis, a copy number of ~120 was estimated for the con-198 sequence. A copy number of 5 was estimated for both the G-2 intervening sequence and the G-1 transcribed region. We interpret these data as showing that the con-198 sequence is present at multiple (~100) genomic sites, of which we have characterised two. In G-2, the con-198 sequence forms part of a degenerate 1.2 kb tandem repeat whereas in G-1 it is present as a single copy.

1.8 kb transcript was clearly distinguishable in the G-1 hybridisation. This 1.8 kb transcript hybridised specifically to the G-1 K probe (Fig. 4B, lane 2) from the 3' end of the clone, which lacked the con-198 sequence. When RNA was separated in 10% acrylamide-urea, G-1X, G-2, G-2H and the con-198 sequence hybridised to a set of four relatively low molecular weight transcripts ranging from 240 to 800 nt (Fig. 4C, lane 3–6). In contrast, G-1 and G-1 K showed no such hybridisation (overexposure of the G-1 probe showed faint signal to the four low molecular weight transcripts, data not shown). These data indicate that the con-198 sequence is present on multiple RNAs including four RNAs ranging from 240 to 800 nt, whereas the region downstream of the con-198 sequence in G-1 codes for one transcript of 1.8 kb.

A series of primer extension experiments were performed to more precisely analyse transcription from G-1 and G-2 (see Figs 1 and 5A for location of primers). Primer P-4 (located immediately downstream of the SL sequence in con-198) yielded a number of extension products, indicating several transcripts or con-198 sequences (Fig. 5B, lane 1). A non con-198 G-2 primer, 6(1), located 520 nt upstream of the 3' con-198 in G-2 resulted in a single extension product of 37 nt (Fig. 5B, lane 2). This mapped to 596 nt upstream of the terminal 'TG dinucleotide of the 3' con-198 (see Fig. 1). Note also the faint corresponding extension product of ~800 nt in primer extension using a primer from the con-198 sequence in Figure 5B, lane 1. Partial deoxy sequencing of this extension product confirmed its origin from the G-2 sequence (data not shown). This transcriptional start site in G-2 is located 71 bp downstream of a consensus TATA box with a CAT sequence a further 34 bp upstream. This initiation site is consistent with RT-PCR results when a series of 5' primers from 370 to 900 bp upstream of the 3' con-198 sequence were used in conjunction with an anti-sense primer from the con-198 sequence (see Figs 1 and 5A for primer sites). Primers downstream of the putative transcriptional start site yielded RT-PCR products of the predicted size but a primer upstream of this start site yielded no product (Fig. 5, panel C).
The 3' end of the transcript encoded by G-2 was localized by RT-PCR to a region between 110 and 150 bp downstream of the SL sequence. Primers con-54 and G-2 3306/3285, located respectively adjacent to the SL sequence and at 109 bp 3' of the SL sequence, resulted in RT-PCR products of the appropriate size when used with 5' primers up to 490 bp from the con-198 sequence, (Fig. 5C, panels 1 and 2). In contrast, an anti-sense primer located a further 40 bp downstream yielded no amplification products when used with the same 5' primers (Fig. 5C, panel 3). RT-PCR using primers complementary to the opposite strand (complementary to 3' primer G-2 3306/3285; and 5' primer G-2 2705/2725 and G-2 2755/2775) resulted in no amplified products (data not shown), confirming that transcription proceeds from the SL-containing strand and not in the opposite direction.

The G-1 K fragment that hybridised to a 1.8 kb transcript on Northern blot analysis (Fig. 4) was located at the 3' end of the clone (nucleotides 2940-3126). Thus primers from this region were used in primer extension analysis. A primer located from nucleotides 3630-3610 (see Figs 1 and 5 A) resulted in a single extension product of 75 nt (Fig. 5B, lane 3). This mapped the transcriptional start site to -3150 nucleotides downstream of the con-198 sequence at a position 167 bp from a consensus TATA sequence and 207 bp from a CAT box.

**DISCUSSION**

Here we describe a 198 bp repetitive element in the *A. cantonensis* genome which contains the 22 nt SL sequence that is normally encoded by the multi-copy SL gene and trans-spliced onto a subset of mRNAs.

This repetitive element, designated con-198, is found in multiple copies (>100) that are contiguous with at least two unlinked DNA sequences, described herein as clones G-1 and G-2. The non con-198 regions of G-1 and G-2 are present in lower copy number (~5), thus G-1 and G-2 may not represent the major genomic context of the con-198 sequence. However, in the
context exemplified by G-2, the con-198 sequence forms part of a 1.2 kb tandem repeat, whereas in G-1 it is present as a single copy. In several features the G-2 sequence encompassing the two con-198 sequences resembles a retrotransposable element (Fig. 1). However, the absence of ORFs coding for elements homologous to retroviral gag, pol or env genes, the low copy number of the con-198 sequence in tandemly reiterated form, and the similarity to a degenerate tandem repeat do not allow us to designate con-198 the LTR of a retrotransposable element.

Rather, we interpret the Northern blot analysis (Fig. 4) as indicating that the con-198 is found on multiple RNAs including a set of four RNAs ranging from 240 to 800 nt. Thus, primer extension analysis using a primer from the con-198 sequence gave a number of extension products corresponding to RNA transcripts from a number of DNA sequences with the con-198 sequence. Primer extension using a primer upstream of the con-198 sequence in G-2 gave one extension product corresponding to transcription from one DNA region with the con-198 sequence, mapping the start site to 726 nt upstream of the SL sequence in con-198. Our results from RT-PCR are consistent with this interpretation, mapping the 5' end to between 680 and 970 nt upstream of the SL sequence in G-2, and mapping the 3' end to between 110 and 150 nt downstream of the SL sequence. Thus transcription from the G-2 genomic sequence results in an 858-888 nt transcript with an internal copy of the SL sequence, and consensus TATA and CAT boxes 73 and 106 nt upstream of the proposed start site. By comparison, in the DNA exemplified by G-1, transcription starts ~3 kb downstream of the con-198 sequence. The biological significance of multiple transcripts with an internal copy of the repetitive element designated con-198 which includes the 22 nt SL sequence is unclear. Similarly, whether or not the presence of the con-198 sequence in any way affects downstream transcription is also unclear.

The curious feature of con-198 is the presence of the SL sequence which is normally trans-spliced onto the 5' end of a subset of nematode mRNAs. Comparison of these A.cantonensis elements containing orphorn SL sequences with the O.volvulus and B.malayi DNA elements containing an orphan SL sequence shows 70% identity in the 65-100 bp flanking the SL sequence (4,20). Thus analogous elements may be found in other nematode species.

In kinetoplastid protozoa trans-splicing is thought to specify the 5' end during the processing of polycistronic precursors into monocistronic mature mRNAs (24). Recent evidence indicates that it may act similarly during the processing of polycistronic RNAs in C.elegans when SL1 trans-splices to the first (5') transcript and SL1 or SL2 trans-splices to 3' RNAs (25,26). It has further been suggested that the leader sequence functions as an efficient translation initiation site (27), and other functions have been proposed (20,28). In A.cantonensis the genomic organisation of orphan SL sequences within a repetitive element and the transcription of these genomic elements either yielding transcripts with an integral SL sequence or transcribing downstream of the SL sequence suggests a biological role(s) other than trans-splicing. This role(s) may be investigated by characterising transcripts with an integral SL sequence; determining whether transcripts are translated and determining whether deletion of the SL and/or con-198 sequence affects transcription or translation.

ACKNOWLEDGEMENTS

We thank Liao Fang-Ming and Yeh Chen-Hsiung for excellent technical assistance, Ms Lola Wen for typing and Dr C. Fletcher for reviewing the manuscript. This work was supported by National Science Council, Taiwan (NSC-81-0412-B001-04).

REFERENCES