The evolution of spliced leader trans-splicing in nematodes

Jonathan Pettitt, Neale Harrison, Ian Stansfield, Bernadette Connolly and Berndt Müller

School of Medical Sciences, Institute of Medical Sciences, University of Aberdeen, Aberdeen AB25 2ZD, Scotland, U.K.

Abstract

Spliced leader trans-splicing occurs in many primitive eukaryotes including nematodes. Most of our knowledge of trans-splicing in nematodes stems from the model organism Caenorhabditis elegans and relatives, and from work with Ascaris. Our investigation of spliced leader trans-splicing in distantly related Dorylaimia nematodes indicates that spliced-leader trans-splicing arose before the nematode phylum and suggests that the spliced leader RNA gene complements in extant nematodes have evolved from a common ancestor with a diverse set of spliced leader RNA genes.

Introduction to SL (spliced leader) trans-splicing

SL trans-splicing is the transfer of a short RNA sequence called SL RNA to the 5′ end of mRNAs by a mechanism involving components of the splicing machinery. SL trans-splicing was discovered more than 20 years ago in trypanosomes [1], and subsequently in other Euglenozoa. SL trans-splicing is also observed in dinoflagellates [2] and some animals, but is absent from fungi and plants. In animals, SL trans-splicing is found in nematodes, flatworms, rotifiers, ciliarians, non-vertebrate chordates, some arthropods (amphipods and copepod crustaceans), ctenophores and hexactinellid sponges [3–8]. It is not clear whether the sporadic phylogenetic distribution of SL trans-splicing reflects multiple parallel losses occurring following descent from a common ancestor that was capable of SL trans-splicing, or multiple independent gains in the lineages in which it is found [9,10].

In SL trans-splicing, the SL sequence is joined to the 5′ end of protein-coding mRNAs (Figure 1A). The SL, which ends with a 5′ splice site, is initially expressed as the 5′ end of a longer precursor RNA (SL RNA) and behaves like an exon in the cis-splicing reaction. The 3′ end of the SL RNA behaves intron-like and contains an Sm protein complex-binding site (reviewed in [11,12]). The second partner in the trans-splicing reaction is the pre-mRNA with an intron-like sequence at the 5′ end that ends with a 3′ splice site (but lacks an upstream 5′ splice site). This intron-like sequence is also referred to as an ‘outron’ [13]. The SL replaces the outron of the target pre-mRNA by a process closely related to the removal of introns during cis-splicing. The first step of the trans-splicing reaction produces a free SL exon ending with a 3′ hydroxy group and a Y-branched structure with the SL RNA intron-like sequence linked to a branch point A in the outron by a 2′,5′-phosphodiester bond [14]. This structure is similar to the lariat structure formed during cis-splicing [11]. In the second step, the 3′ hydroxy group of the SL attacks the 3′ splice site. This leads to the joining of the two exons to form the mRNA 5′ end, and the release of the branched intron RNA.

SL RNAs are capped with modified cap structures such as the trimethylguanosine cap structure in nematodes [15] or a 7-methylguanosine cap and additional modifications on the first four nucleotides (the ‘cap 4’ structure) in trypanosomes [16]. The trans-splicing reaction results in the transfer of the cap structure, together with the SL sequence, to the 5′ end of the resulting mRNA. The cap structure is important for mRNA translation, being instrumental for the recruitment of ribosomes during translation initiation. The SL may also have additional roles in translation; for example, in flatworms, an AUG in the SL sequence can serve as translation initiation codon [17], and, in nematodes, the SL stimulates translation of trimethylguanosine-capped mRNA [18,19]. Other SL functions probably include mRNA stabilization and the removal of 5′-untranslated regions that may contain detrimental sequences [20,21].

Among the SL RNAs identified to date, there is little or no significant sequence conservation between the various phylogenetic groups. The SL RNAs are normally up to 150 nt long, and the length of the SL varies greatly. SL RNAs from kinetoplasts and nematodes can be folded into a characteristic structure containing three stem–loops (Figures 1B and 2B), whereas others, such as from the chordate Ciona intestinalis, are not predicted to adopt such a structure [6,9]. However, one conserved feature is the presence of a short Sm motif sequence in the SL RNA intron. In spliceosomal snRNPs and the U7 snRNP involved in processing of histone mRNA, the Sm motif is crucial for assembly of a functional RNA–protein particle and the methylation of the cap structure [22,23]. This indicates that, in the organisms surveyed, the SL RNAs are related to, or possibly derived from, snRNAs involved in mRNA splicing.
**Figure 1** | SL trans-splicing

(A) The trans-splicing reaction. Trans-splicing occurs between SL RNA composed of the SL exon sequence ending with a 5′ splice site (5′ ss) (exon 1) and the 3′ intron-like portion of the RNA, and mRNA with intronic sequences including a 3′ splice site (3′ ss) (outron) at the 5′ end followed by an exon sequence (exon 2) [11,12]. Intermediates of the trans-splicing reaction are a free SL and a Y-branched intermediate composed of the SL RNA intron-like portion linked to the branch point A in the outron. The two exons are joined by the attack of the 3′ hydroxy group at the SL 3′ end on to the 3′ splice site, and the Y-branched intronic RNA is released. (B) Nematode SL RNA structure indicating stem-loops 1–3 and Sm motif. The SL sequence is shown as a thicker line. The consensus sequence of *C. elegans* SL1 is highly conserved, whereas there is some variability between SL2 RNA sequences. (C) SL1 RNA trans-splicing: SL1's trans-splice to RNAs with 5′ ends containing intronic sequences ('outrons') including a U-rich outron element (OUE) [42]. (D) Operon resolution by SL2. SL2 RNA is recruited to operons with ~100 nt intergenic regions containing a U-rich downstream sequence element (U-rich DSE) and a separate U-rich element (Ur element) closer to the 3′ splice site [42]. SL2 trans-splicing is dependent on 3′ end formation of the upstream gene; it requires a functional polyadenylation signal and an interaction between the polyadenylation factor CstF64 and SL2 RNA [44,45,59]. The U-rich DSE is thought to be the binding site for CstF. (E) Operon resolution by SL1. SL1 acts in the resolution of operons with minimal distance between polyadenylation signal and 3′ splice site [37].

---

**Nematode phylogeny**

Recent molecular phylogenetic analyses of the nematode phylum provide strong support for the existence of multiple clades of nematodes [24–26] (Figure 3). However, most of the molecular information concerning nematode biology has come from a single clade, Rhabditina (Clade V), containing *Caenorhabditis elegans*. Two other clades, Spirurina (Clade III) and Tylenchina (Clade IV) which include important animal and plant parasites have also contributed to our understanding of nematode molecular biology. There is strong support for these three clades forming a monophyletic group, the Rhabditida. Therefore molecular studies of these nematodes do not necessarily give an insight into the molecular biology of nematodes in general. Indeed, there is evidence to suggest that the Rhabditida nematodes share many derived features not present in the ancestor of the extant nematodes.

Until recently, molecular information from the other nematode clades was limited, but the availability of the draft
Figure 2 | *T. spiralis* and *T. pseudospiralis* SL RNAs

(A) Alignment of SL sequences from *T. spiralis* (TSLs) with *C. elegans* SL1 and SL2. SLs also found in *T. pseudospiralis* are indicated. Conserved sequence features with similar positioning within SLs are highlighted. (B) TSL2 RNA secondary structure. Secondary-structure prediction was carried out using Mfold [60]. Stem-loops 1, 2 and 3 are indicated, the SL is outlined and the Sm motif is underlined.

| TSL1* | AGGAUAACCCGGAAUACCAUAGGC | 22 |
| TSL2 | AGGAUAACCCGGAAUACCAUAGGC | 22 |
| TSL3 | GGUAAUACCCGGAAUACCAUAGGC | 22 |
| TSL4 | GCGAUAACCCGGAAUACCAUAGGC | 24 |
| TSL5 | AACCCGGAAUACCAUAGGC | 22 |
| TSL6 | AACCCGGAAUACCAUAGGC | 22 |
| TSL7* | AACCCGGAAUACCAUAGGC | 22 |
| TSL8* | AACCCGGAAUACCAUAGGC | 22 |
| TSL9 | AACCCGGAAUACCAUAGGC | 22 |
| TSL10 | AACCCGGAAUACCAUAGGC | 22 |
| TSL11* | AACCCGGAAUACCAUAGGC | 22 |
| TSL12 | AACCCGGAAUACCAUAGGC | 22 |
| TSL13 | AACCCGGAAUACCAUAGGC | 22 |
| TSL14 | AACCCGGAAUACCAUAGGC | 22 |

* SLs also found in *T. pseudospiralis*

Figure 3 | Nematode phylogeny and SL RNA gene evolution

This simplified nematode phylogeny is based on small ribosomal subunit RNA sequences [24,26,61]. The studies are agreed on the relative position of *C. elegans* with respect to the basal nematodes *T. spiralis* and *T. pseudospiralis* and *P. punctatus*, but the exact topology differs. Additional less-well-positioned clades were omitted for clarity. We propose that the last common nematode ancestor had a diverse set of SL RNA genes. A combination of sequence diversification and loss of SL RNA genes that were present in the last common nematode ancestor results in the SL RNA gene spectrum found in extant nematodes.

genome sequence of *Trichinella spiralis* has provided the first insight into the molecular biology of a nematode that falls outside the Rhabditida grouping [27,28]. *T. spiralis* is a member of the Dorylaimia clade (Clade I), which along with the Enoplia (Clade II), occupies an important phylogenetic position at the base of the nematode phylogeny [24,25]. Consistent with this, these clades possess several molecular, cellular and embryological traits that are found in other animal groups, but which are absent from the Clade III/IV/V nematodes [29–33]. Thus studying these ‘basal’ nematode clades is essential for a better understanding of nematode biology in general.

**SL trans-splicing in Rhabditida nematodes**

SL trans-splicing in nematodes was first observed in *C. elegans* [34], and our view of this process in nematodes is to a large part determined by work conducted in this organism, and by information from the gut parasite *Ascaris suum*, another member of the Rhabditida.

SL trans-splicing in *C. elegans* is the subject of several excellent reviews [35–37]. Approx. 70% of *C. elegans* mRNAs are subject to SL trans-splicing to one of two SLs, SL1 or SL2, which are transcribed as ∼100–110 nt long SL RNAs (Figure 1B) [38–40]. In *C. elegans*, the ∼110 SL1
RNA genes occur in the same tandem repeat fragment as 5S rRNA on chromosome V [41]. The genes for the SL2 RNAs are fewer and more dispersed [41]. Some 19 SL2 RNA gene variants, which show more sequence variation than SL1 RNAs, are listed in WormBase (http://wormbase.org/).

SL1 is the predominant SL and is primarily trans-spliced to monocistronic pre-mRNAs (Figure 1C). A bioinformatic analysis of C. elegans outon sequences involved in trans-splicing has led to the identification of a conserved UC-rich outon element (‘Ou’ element) located ∼50 nt upstream of the splice site [42].

C. elegans contains at least 1000 operons composed of two to eight genes, representing approx. 15% of all C. elegans genes [43]. The vast majority of the multicistronic transcripts from these operons are resolved into monocistronic mRNAs in a process dependent on SL2 RNA (Figure 1D). Typically, operons resolved by SL2 have a ∼100 nt region between the polyadenylation signal and the 3′ splice site which contains a U-rich (Ur) element [44,45]. A bioinformatic investigation of such operons has led to the identification of two Ur elements: a Ur putative CstF (cleavage-stimulation factor)-binding site (the Ur downstream element located just 3′ of the polyadenylation site), and the Ur element normally 40–60 nt downstream of the polyadenylation site [42,45]. SL2 RNA is known to interact with CstF [44], a factor involved in mRNA cleavage at the polyadenylation signal. Binding of CstF to the Ur downstream element would facilitate the recruitment of SL2 RNA, thus coupling RNA 3′ end formation with trans-splicing. In a few cases, the intergenic region in operons is much shorter and the polyadenylation signal is followed very closely by the 3′ splice site. Such transcripts are resolved by trans-splicing of SL1 (instead of SL2) to the 3′ splice site (Figure 1E) [37].

A comparison of factor requirements for in vitro trans-splicing using Ascaris extracts revealed that, similar to cis-splicing, trans-splicing requires U2, U4, U6 and U5 snRNA [46,47]. U1 snRNA, however, is not required [46]. Interestingly, attraction of the SL sequence to a U1 snRNA fragment is sufficient for this RNA to participate in trans-splicing [48], indicating that SL RNAs and spliceosomal snRNAs are related. Similar to cis-splicing, SR proteins are required for trans-splicing, and participate in the recruitment of U2 snRNP to the branch point region [49–51], presumably by interaction with splice enhancer domains in the exonic sequence [9,50].

Like snRNAs, SL RNAs also associate with Sm/Lsm proteins through the Sm motif and have a trimethylguanosine cap [15,48,52]. In addition, two SL RNA–associated proteins with apparent molecular masses of 175 and 30 kDa specifically required for SL trans-splicing were identified in Ascaris [52,53]. The C. elegans SL75P and SL21P are homologues of the Ascaris proteins and associate specifically with SL1 snRNA [54].

A recent survey of operons and SL trans-splicing in nematodes addressed whether the emergence of SL2 RNA may be linked to the formation of operons [55]. Conserved operons were found in Rhabditida nematodes (C. elegans, Pristionchus pacificus, Nippostrongylus brasiliensis, Strongyloides ratti, Brugia malayi and Ascaris sp.) [55]. An analysis of transcripts from these operons revealed that, in Rhabditina, the SL2 or SL2-like RNAs are used only for the resolution of operons. In the spirurine nematodes, B. malayi and A. suum, operonic transcripts are resolved using SL1 and in the tylenchine nematode, S. ratti, they are resolved using SL1-like variants. This suggests that the evolution of SL2 RNAs was not linked to the emergence of operons [55] and led to the proposal that they may have evolved specifically in the Rhabditina clade. However, SL2-like SLs have recently been identified in the tylenchine nematode Aphelenchus avenae [56], and by us in a dorylaimid nematode [57], indicating that SL2 is likely to be more ancient than previously proposed.

### SL trans-splicing in dorylaimid nematodes

We have analysed trans-splicing in the dorylaimid nematodes T. spiralis and Trichinella pseudospiralis, both intramuscular parasites infecting mammals and birds respectively. This has led to the identification of a set of unusual highly polymorphic SL RNAs [58]. During our initial investigation of gene expression in T. spiralis using the 5′ RACE (rapid amplification of cDNA ends) technique, we observed that several unrelated cDNAs had identical 21–24 nt sequences at the very 5′ end. An interrogation of the T. spiralis draft genome indicated that these sequences were not contiguous with the genomic sequences from which the rest of the cDNA was derived, and we identified these as T. spiralis SL sequences (TSLs). Interestingly, the TSLs have only limited similarity to each other and no significant homology with C. elegans SL1 or SL2 except for the last two nucleotides (AG), which form part of the 5′ splice site, and two regions containing one or two Us. One of these regions is located just 5′ of the splice site, and the second one is located closer to the 5′ end (Figure 2A).

We were able to clone several full-length T. spiralis SL RNAs containing the SL sequence joined to the 68–80 nt intron-like sequence [58]. These molecules can adopt the secondary structure, with three stem–loops, typical for nematode SL RNAs: the SL sequence is part of stem–loop 1; and stem–loops 2 and 3 flank the Sm motif (Figure 2B). Identical and similar SL sequences were also observed in the sister species T. pseudospiralis (Figure 2A).

A bioinformatic search for SL RNA revealed that the SL RNA genes, which we named tslr genes for trans-spliced leader RNA, are found together on a contiguous region of the T. spiralis genome, indicating that they are located together as in other organisms [58]. However, unlike in C. elegans and its close relatives, the tslr genes are not organized in tandem clusters with the 5S rRNA genes, although these latter rRNA genes are also found in the same region of the genome. Instead, the tslr genes are distributed over an 8 Mb region interspersed with putative protein-coding genes. As the T. spiralis genome sequence is not complete, it is possible that we have not yet identified all of the tslr genes.
The identification of these diverse SL RNAs in the two *Trichinella* species is surprising. We have so far no evidence for the existence of operons in *Trichinella*, thus it is at this stage unlikely that the *Trichinella* SL RNAs fall into different groups acting on monocistronic and polycistronic transcripts as in rhabditine nematodes. In addition, we have so far not been able to identify SL1 or SL2-like SL RNAs in *Trichinella*. Interestingly, no homologues of the proteins associated with SL1 RNA have been identified to date in *T. spiralis*. However, our recent observation of an SL2-like SL RNA in a Clade I nematode, *Prionchulus punctatus*, suggests that the SL RNA repertoire in Clade I nematodes is greater than that which can be inferred directly from *Trichinella* species [57].

Together, our findings and those of others indicate that SL RNA *trans*-splicing was part of the gene expression mechanism of the last common nematode ancestor. We propose that current nematode SL RNA gene complements may have arisen from an original set of SL RNAs including the ancestors of SL1 and SL2 SL RNA genes in the last common nematode ancestor by differential loss of SL RNA genes combined with diversification of SL RNA sequences (Figure 3). In the Rhabditida, all but SL1 and SL1-like SL RNA genes and, in some cases, SL2 or SL2-like SL RNA genes were lost. In the Dorylaimia, non-SL1/SL2 SL RNA genes were maintained. The highly polymorphic SL RNA in the *Trichinella* species would represent an example of sequence variation of non-SL1/SL2 SL RNA genes. The presence of SL2-like SL RNA genes in a dorylaimid nematode indicates that these genes were not necessarily lost during the evolution of Clade I nematodes.

**Conclusions**

Since SL *trans*-splicing has been detected in members of widely divergent nematode clades (the last common ancestor of *C. elegans* and *T. spiralis* was likely to be the last common ancestor of all known extant nematodes), it is reasonable to conclude that it was present in the ancestral nematode lineage. The biological role of SL RNA *trans*-splicing has been well studied in *C. elegans* and a few additional Rhabditida nematodes. However, an important question is how applicable these findings are to the more ‘basal’ dorylaimid clades. The qualitative and quantitative role of SL *trans*-splicing might well vary between the various nematode groups. For instance, the diversity of SL RNAs observed in *Trichinella* species might well reflect functional diversification. Addressing this issue will require the development of tools that allow the investigation of biological problems in these nematodes.

In addition, since it seems highly likely that SL *trans*-splicing is widespread throughout this medically and economically significant phylum, then better understanding of the molecular basis of this process, especially identifying components conserved throughout the phylum, offers the opportunity of developing specific therapeutic interventions against both human, animal and plant parasites.

**Funding**

This work was support by the Wellcome Trust [grant number 076220].

**References**

15. Thomas, J.D., Conrad, R.C. and Blumenthal, T. (1988) The *C. elegans* *trans*-spliced leader RNA is bound to Sm and has a trimethylguanosine cap. Cell 54, 533–539