Operons are a conserved feature of nematode genomes

Jonathan Pettitt, Lucas Philippe, Debjani Sarkar, Christopher Johnston, Henrike Johanna Gothe, Diane Massie, Bernadette Connolly and Berndt Müller

School of Medical Sciences, University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen AB25 2ZD, UK

*T. spiralis* and *T. muris* mRNA 5’ ends are deposited under GenBank accession numbers KF442418-KF442435; KF511776; KF511777; KF768019.
Running Title: Conservation of nematode operons

Keywords: operons; SL trans-splicing; nematode genome

Corresponding Author:

Jonathan Pettitt

School of Medical Sciences

Institute of Medical Sciences

Aberdeen AB25 2ZD

UK

Tel.: +44 (0)1224 437516

FAX: +44 (0)1224 437465

Email: j.pettitt@abdn.ac.uk
ABSTRACT

The organisation of genes into operons, clusters of genes that are co-transcribed to produce polycistronic pre-mRNAs, is a trait found in a wide-range of eukaryotic groups, including multiple animal phyla. Operons are present in the class Chromadorea, one of the two main nematode classes, but their distribution in the other class, the Enoplea, is not known. We have surveyed the genomes of *Trichinella spiralis*, *Trichuris muris* and *Romanomermis culicivorax* and identified the first putative operons in members of the Enoplea. Consistent with the mechanism of polycistronic RNA resolution in other nematodes, the mRNAs produced by genes downstream of the first gene in the *T. spiralis* and *T. muris* operons are trans-spliced to spliced leader RNAs, and we are able to detect polycistronic RNAs derived from these operons. Importantly, a putative intercistronic region from one of these potential enoplean operons confers polycistronic processing activity when expressed as part of a chimeric operon in *C. elegans*. We find that *T. spiralis* genes located in operons have an increased likelihood of having operonic *C. elegans* homologues. However, operon structure in terms of synteny and gene content is not tightly conserved between the two taxa, consistent with models of operon evolution. We have nevertheless identified putative operons conserved between Enoplea and Chromadorea. Our data suggest that operons and SL trans-splicing predate the radiation of the nematode phylum, an inference which is supported by the phylogenetic profile of proteins known to be involved in nematode SL trans-splicing.
INTRODUCTION

The organisation of open-reading frames into operons, such that multiple, distinct gene products are produced from a single, polycistronic transcript, is commonplace in prokaryote genomes (Jacob et al. 2005). Operons are also found in eukaryotes, although their distribution is sporadic and it does not seem likely that they represent an ancestral eukaryotic trait (Lawrence 1999; Hastings 2005). In prokaryotes, translation of multiple open reading frames in a polycistronic RNA occurs through multiple independent translation initiations. In eukaryotes, the polycistronic RNAs must first be processed into individual mRNAs before being translated. This creates a problem in that the processed, downstream mRNAs would lack a cap structure necessary for RNA stability and translation. A number of eukaryotes are able to circumvent this problem through the trans-splicing of a short “spliced leader” (SL) RNA onto the 5’ end of the mRNA. Because the precursor SL RNAs that donate the SL are trimethylguanosine-capped the trans-splicing event provides the cap structure for the mRNA. Thus by providing a mechanism that allows the formation of monocistronic, capped mRNAs from polycistronic RNA, SL trans-splicing enables the organisation of eukaryotic genes into operons. It is striking that, at least to date, all eukaryotes in which operon usage is widespread also undergo SL trans-splicing (Johnson et al. 1987; Spieth et al. 1993; Davis and Hodgson 1997; Blumenthal et al. 2002; Gannot et al. 2004; Guiliano and Blaxter 2006; Satou et al. 2008; Marlétaz et al. 2008; Dana et al. 2012; Protasio et al. 2012; Tsai et al. 2013), suggesting that the resolution of polycistronic RNA is dependent upon SL trans-splicing.

Although operon organisation is widespread in numerous eukaryotic taxa, the evolutionary mechanisms that have resulted in this form of gene organisation are not well understood. The most detailed analysis of the problem has come from studies in C. elegans, which led to the hypothesis that operon organisation allows the marshalling of multiple genes
under the control of a single promoter. This makes cells better able to cope with situations when transcription factors are present in limiting concentrations, such as recovery from growth arrest (Zaslaver et al. 2011). However, it is far from clear whether this is the only mechanism responsible for the evolution of operon organisation, and the general applicability of this hypothesis to other members of the nematode phylum is not known.

In order to better understand the relationship between operon evolution and SL trans-splicing it is necessary to determine the distribution of operon organisation across the nematode phylum. Nematodes can be divided into two major classes: Enoplea and Chromadorea (Holtermann et al. 2006; Meldal et al. 2007), with the latter class being much better characterised in terms of gene expression mechanisms, largely because it contains C. elegans. Both SL trans-splicing and operons have been identified in multiple nematodes within the Chromadorea (Evans et al. 1997; Lee and Sommer 2003; Guiliano and Blaxter 2006). However, the presence of operons has not been reported in nematodes from the other taxon.

We have previously identified SL trans-splicing in the enoplean nematodes Trichinella spiralis and Prionchulus punctatus (Pettitt et al. 2008; Harrison et al. 2010), suggesting that they may also possess operons. The draft genome of T. spiralis should be a useful resource for identifying operons in this nematode (Mitreva et al. 2011). However, identification of operons is not straightforward. The original discovery of operons in C. elegans was dependent upon the discovery of a specific spliced-leader, SL2, which is trans-spliced to most mRNAs derived from genes downstream of the first gene in operons (Spieth et al. 1993), but not all nematodes use a specialised SL RNA to resolve polycistronic RNAs (Guiliano and Blaxter 2006). Thus, this feature cannot be considered diagnostic for mRNAs derived from nematode operons. The other feature common to operonic genes is that, at least in C. elegans, the
distance between genes in an operon (the intercistronic region, or ICR) is unusually short, with a mean ICR size of 126 bp (Blumenthal et al. 2002). Again, this trait is not definitive: the ICR size can be considerably larger in the operons of other nematodes (Guiliano and Blaxter 2006; Ghedin et al. 2007) and even in C. elegans operons exist with large ICR distances (Morton and Blumenthal 2011).

Previous approaches to identify operons in _T. spiralis_ (Mitreva et al. 2011) looked for pairs of _T. spiralis_ genes whose homologues were in the same operon in _C. elegans_. This resulted in a limited set of 16 neighbouring pairs of genes that potentially correspond to _T. spiralis_ operons, however further characterisation of these candidate operons was not undertaken. We have used conserved synteny, coupled with the fact that mRNAs derived from downstream genes in operons are dependent on SL trans-splicing to elucidate a set of putative _T. spiralis_ operons. Detailed analysis of two of these putative operons indicates that they display all the molecular characteristics expected of loci that generate polycistrionic RNA. Taken together our data indicates that the organisation of genes into operons was present in the last common ancestor of the Chromadoria and Enoplea.
MATERIALS AND METHODS

Bioinformatic identification of T. muris SL RNA genes

T. muris SL RNA genes Tmu-SL1, Tmu-SL2, Tmu-SL3 and Tmu-SL9 were identified by searching the T. muris genome dataset with T. spiralis SL sequences using the BLASTN tool. Hits were considered if two of the three following criteria were met: a candidate Sm protein binding site was detected (AATTTTTG), the 5’ splice site sequence was conserved (AGGT) and a run of at least three T’s was found located approximately 100 base pairs from the end of the putative SL sequence. T. muris SL RNA genes Tmu-SL4, Tmu-SL5, Tmu-SL6, Tmu-SL7, Tmu-SL8 and Tmu-SL10 were identified by searching the T. muris genome dataset with the Tmu-SL1, Tmu-SL2, Tmu-SL3 or Tmu-SL9 sequences using the BLASTN tool, and fulfilling the same criteria as above. Genes for Tmu-SL1, Tmu-SL2, Tmu-SL6, Tmu-SL8, Tmu-SL9, Tmu-SL10 and Tmu-SL11 were also identified with a PERL script (Pettitt et al, 2008) used previously to identify T. spiralis SL genes, except that the parameters for the Sm binding site were changed to AATTTTTG/TG.

Analysis of T. spiralis SL containing ESTs and Identification of Putative Conserved Operons

The ESTs were identified earlier (Pettitt 2008). To identify the corresponding gene from which each EST was derived, the EST sequences were mapped onto the T. spiralis draft genome sequencing using BLASTN. The corresponding gene was annotated as being in a putative operon if its upstream or downstream neighbour genes were on the same DNA strand with an intercistronic distance of ≤ 1 kb. If the neighbours were on the same strand, but between 1 kb and 5 kb away, they were recorded as ambiguous. Otherwise the genes were annotated as non-operonic. A minority of the ESTs matched to more than one predicted T. spiralis gene.
To identify the operonic status of the *C. elegans* homologues of each SL trans-spliced *T. spiralis* EST, BLASTX searches were carried out. We used an E-value cutoff of $10^{-5}$ to determine homology. In addition, if we obtained similar E-values for multiple *C. elegans* genes we excluded that EST from the analysis.

*T. spiralis* and *R. culicivorax* homologues of operonic gene pairs conserved between *C. elegans* and *B. malayi* (Ghedin et al. 2007) were identified using BLASTP searches with the *C. elegans* upstream homologues from each pair as a query in searches against the respective gene predictions. The predicted coding region of the *T. spiralis/R. culicivorax* gene immediately downstream of the gene identified by this search was then used as a query sequence in a “reciprocal” BLASTP search against the *C. elegans* gene predictions. Since it was apparent that *T. spiralis* genes, which are separated by unusually short intergenic distance (such as might be expected to genes organised into operons), are prone to misannotation and conflation into a single gene prediction, we also carried out an additional step using both *C. elegans* gene pairs as query sequences in BLASTP searches against the *T. spiralis* gene predictions, looking for cases where both *C. elegans* genes return matches to the same *T. spiralis* gene. Manual examination of such putative gene prediction errors, guided by the results of the sequence similarity searches, was then used to identify the intercistronic regions in each case. In all BLASTP searches we used an E value cut-off of $10^{-5}$ for the establishment of homology.

**Phylogenetic Profiling of SL trans-splicing snRNPs**

Homologues of *C. elegans* *sna-1, sut-1* and *sna-2* were identified by carrying out BLAST searches against the NCBI non-redundant database, except in the case of *R. culicivorax* and *T.
*muris*, where BLAST searches were carried out against datasets downloaded from
http://www.nematodes.org/genomes/romanomermis_culicivorax/ and
http://www.sanger.ac.uk/resources/downloads/helminths/trichuris-muris.html,
respectively. Phylogenetic tree construction was carried out with the online implementation
of PhyML (Dereeper et al. 2008) using default settings.

**Nematode isolation and RNA preparation**

*T. spiralis* RNA was produced as described (Pettitt et al 2008). *T. muris* RNA was a generous
gift from Allison Bancroft and Richard Grencis, University of Manchester

**Analysis of RNA 5’ ends**

The 5’ ends of cDNAs were obtained through 5’ RACE using the GeneRacer kit (Invitrogen),
according to the manufacturer’s instructions. Gene-specific primers used are given in Table S3,
and the cDNAs, amplified by PCR, using either GoTaq polymerase (Promega), or Expand High
Fidelity polymerase (Roche), were cloned into pGEM T-Easy (Promega). The resulting plasmid
inserts were sequenced by the University of Dundee Sequencing Service.

**Detection of processing intermediates of polycistronic transcripts.**

RNA was reverse transcribed using SuperScript™ III Reverse Transcriptase (Invitrogen) and
random primers according to the instructions of the manufacturer. In control reactions (-RT)
all reagents were included except the reverse transcriptase. Processing intermediates were
normally amplified by two rounds of PCR with nested primer pairs (Table S4) and either
GoTaq polymerase (Promega), or Expand High Fidelity polymerase (Roche) and visualised by
agarose gel electrophoresis. The identity of the PCR products was determined by cloning into
pGEM Teasy (Promega) and sequencing of plasmid inserts.

**Identification of SL RNA 3’ ends.**
SL RNA 3’ ends were determined essentially as described previously (Pettitt et al 2008). *T. muris* total RNA (~5 µg) was poly(A) tailed using yeast poly(A) polymerase, reverse transcribed using an oligo-dT-anchor primer (GCGAGCTCCGCGCCGCGCTTTTTTTTTTTTTTTT), and then PCR amplified using an SL-specific primers (GGTTAATTACCAATTTAAAAG) and an anchor primer (GCGAGCTCCGCGCCGCG). PCR fragments were inserted into pGEM-T easy (Promega), and inserts were sequenced at the University of Dundee DNA Sequencing Facility.

**SL RNA secondary structure prediction**

Secondary structure prediction of *T. muris* SL RNA was performed using MFOLD Version 2.3 (Zuker 2003) using the default folding conditions (1 M NaCl, 37°C), and with the constraint that the Sm-binding site (5’-AAUUUUGUUG-3’) was required to be single stranded.

**Generation and analysis of synthetic operon constructs**

The GFP coding region was amplified from pTG96 using the primers 5’-

CAATACTACAGCTCCGGAGTTGCGCAAAGGACTCCAAA-3’ and 5’-

GCTCACATGCTAG CCTATATGTGATATGCATGC-3’. The mCherry coding region, coupled to the *unc-54* 3’ UTR, was amplified from pPD95.75Cherry (a derivative of pPD95.75 in which the GFP coding region was replaced by mCherry) using the primers 5’-

ACAAATAGGCTAGCATGGTGAGCAAGGGCGAG-3’ and 5’-

CGCGGAGACAGGAGGGCCAGGAAAAGAGTTTGTGTTGAT-3’. The primers were designed so that they had overlapping complementary 5’ extensions that introduced an *NheI* restriction site. The two amplicons were purified and fused using a PCR fusion strategy (Hobert 2002). The resulting amplicon consisting of the GFP and mCherry coding regions flanking an *NheI* site was cloned into *Smal-Apal* cut pTG96 using In-Fusion HD (Clontech Laboratories, Inc.) to generate pTG96-Op. The ICRs were cloned from PCR products amplified from genomic DNA. The Tsp-cpt-2-nuaf-3 ICR was amplified from *T. spiralis* genomic DNA using primers 5’-
ATACAAATAGGCTAGCAATTATCCTTTTATAAC-3’ and 5’-
TGCTCACCATGCTAGCTTAGCCAAACTAGGAAATTATTGA-3’ and the Cel-cpt-2~prx-14 ICR was amplified from C. elegans genomic DNA using primers 5’-
ATACAAATAGGCTAGCTTTGTATGACATTATTTAAT3’ and 5’-
TGCTCACCATGCTAGCTTAGCCTACCTAAAAAT-3’, and the Cel-cpt-2~prx-14 ICR was amplified from C. elegans genomic DNA using primers 5’-
ATACAAATAGGCTAGCTTTGTATGACATTATTTAAT3’ and 5’-
TGCTCACCATGCTAGCTTAGCCTACCTAAAAAT-3’.
The resulting PCR products were cloned into NheI cut pTG96-Op using In-Fusion HD (Clontech Laboratories Inc) and the resultant plasmids, pPE#LP1 (Tsp-cpt-2~nuaf-3 ICR clone) and pPE#LP2 (Cel-cpt-2~prx-14 ICR clone) were sequenced to confirm the integrity of the cloning process.

To generate transgenic C. elegans strains, the plasmids were co-injected (100 ng/µl) with Pmyo-2::dTomato (10 ng/µl) into Bristol (N2) wild type hermaphrodites. For each construct several lines were obtained, each of which gave identical expression patterns. Single lines for each construct were selected for the experiments reported here: PE612 - feEx304 [sur-5::gfp::ICRTsp-cpt-2~nuaf-3::mCherry Pmyo-2::dTomato] and PE613 - feEx305 [sur-5::gfp::ICRCel-cpt2~prx-14::mCherry Pmyo-2::dTomato]. Trans-splicing of reporter gene transcripts was analysed as described previously (Harrison et al 2010).

Briefly, total RNA was reverse transcribed and trans-spliced transcripts were PCR amplified using C. elegans SL2-specific (5’ GGTTTTAAACCAGTTACCTCAAG 3’) and mCherry-specific (5’ CCGTCCTCGAAGTTCATCAC 3’) primers. Primers derived from gpd-1 (5’ CCAACTGTCTGCGCACCCTACTCAAC 3’ and 5’ GTCTTCTGCGGTTGCGTACAC 3’) were used to normalise the reactions.

cDNA fragments were cloned into pGEM-T easy (Promega), and inserts were sequenced at the University of Dundee DNA Sequencing Facility.
RESULTS

A putative enoplean operon

As part of the analysis of the transcriptome of the free-living enoplean, Prionchulus punctatus, we identified an EST corresponding to an SL trans-spliced mRNA. Sequence similarity searches using this sequence identified a single predicted T. spiralis gene, Tsp_06075. However, subsequent sequence analysis of Tsp_06075 showed that it corresponds to an erroneous gene prediction, which conflates three genes that are the orthologues of the C. elegans genes zgpa-1 (C33H5.17), dif-1 and aph-1, respectively. That Tsp_06075 is actually three separate genes was confirmed by sequence analysis of 5’RACE products. It seems likely that the unusually short intergenic regions that exist between these three T. spiralis genes caused the gene annotation error (Fig. 1). Such short intergenic distances are characteristic of nematode genes that are arranged into operons (Blumenthal et al. 2002; Guillian and Blaxter 2006; Ghedin et al. 2007), and we thus decided to investigate the possibility that Tsp-zgpa-1, Tsp-dif-1 and Tsp-aph-1 constitute an operon. In parallel we also analysed the homologues of these three genes in the closely related enoplean, Trichuris muris, which show the same syntenic arrangement, although the intergenic distance between Tmu-dif-1 and Tmu-aph-1 is much larger than expected for an ICR. The three C. elegans homologues although organised into operons, are not found in the same operon. However, in a close relative of C. elegans, Pristionchus pacificus, zgpa-1 and dif-1 could potentially constitute a single operon, but again the intergenic space between the two genes is also relatively large compared to the average size of ICRs in C. elegans operons.

We determined the overall pattern of SL trans-splicing of the mRNAs derived from the putative operons in both T. spiralis and T. muris by analysing the 5’ ends of zgpa-1, dif-1 and aph-1 mRNAs using 5’RACE (Fig. 1A)(Table S2). The analysis of Tsp-zgpa-1 and Tmu-zgpa-1
transcripts mapped the mRNA 5’ ends to a region 200-250 base pairs upstream of the start codon, and we failed to detect any SL trans-spliced transcripts derived from this gene in either nematode. In contrast, all dif-1 transcripts analysed were subject to SL trans-splicing in both organisms (Table S2). Tsp-dif-1 transcripts were trans-spliced to Tsp-SL10 (note this SL was previously given the designation TSL-10 (PETTITT et al 2008), but we have renamed it to conform to accepted nematode gene nomenclature rules, which employ a species-specific prefix (BEECH et al. 2010)) and Tmu-dif-1 transcripts were trans-spliced to the newly identified Tmu-SL1, Tmu-SL4 and Tmu-SL12 (Table S2). Analysis of aph-1 transcripts showed that in some cases the transcripts are SL trans-spliced, but we were also able to detect transcripts that initiated approximately 200-300 bp upstream of the start codon, indicating that they were not subject to SL trans-splicing (Fig. 1A; Table S2). It is notable that the distance between Tmu-aph-1 and Tmu-dif-1 is relatively large suggesting the possibility that there are promoter elements immediately upstream of Tmu-aph-1 that would allow the production of transcripts without the need for SL trans-splicing. Such “hybrid operons” have been described in C. elegans (HUANG et al. 2007).

As part of this analysis we identified spliced leaders in T. muris, leading to the discovery of thirteen Tmu-SLs (Figure 2; Table S2). Previous studies have shown that the primary sequences of spliced leaders in T. spiralis are much more variable than those found in the Chromadorea (PETTITT et al. 2008), and many lack the conserved motifs that characterise spliced leaders from these latter nematodes. In contrast, those of P. punctatus do not show the same diversity, and display a greater degree of sequence similarity to the Chromadorid spliced leaders (HARRISON et al. 2010). Analysis of the thirteen distinct T. muris spliced leaders, designated Tmu-SL1-13, support this view, since the T. muris spliced leaders possess the same 5’ GGUWW and central CCC motifs that are highly conserved in the P. punctatus spliced
leaders and Chromadorid SL1 and SL2 families, but missing in most of the *T. spiralis* spliced leaders. The presence of canonical nematode spliced leaders in *T. muris* and *P. punctatus*, despite the fact that the former nematode is more closely related to *T. spiralis*, supports the inference that the *T. spiralis* spliced leaders are derived features.

If *zgpa-1*, *dif-1* and *aph-1* are components of a *bona fide* operon in the two enoplean nematodes, we would expect to be able to detect the polycistrionic RNA from which their mRNAs are derived. Although not a definitive criterion, the presence of polycistrionic, partially processed pre-mRNAs is a predicted property of operon usage. We tested for the presence of such RNA molecules in both *T. spiralis* and *T. muris* (Fig. 1B; Fig. S1) by reverse transcription of total RNA followed by PCR with gene-specific primers. PCR products were then analysed by agarose gel electrophoresis. In *T. spiralis* we detected RNA species connecting the open reading frames of Tsp-*zgpa-1* with Tsp-*dif-1* (*Tsp-zgpa-1~dif-1*), and Tsp-*dif-1* with Tsp-*aph-1* (*Tsp-dif-1~aph-1*) (Fig. 1B). As we failed to amplify any products in control reactions performed in parallel with RNA subjected to mock reactions without reverse transcriptase (Fig. S1), these products represent processing intermediates of polycistrionic transcripts.

The *Tsp-zgpa-1~dif-1* intermediates contained the intercistronic region, and two of the intermediates lacked introns. The *Tsp-dif-1~aph-1* processing intermediates detected were all subject to *cis*-splicing of *dif-1* introns, but we failed to detect an intermediate containing the complete ICR. Instead, the ICR was removed by *cis*-splicing of a cryptic splice donor site located in exon 7 of the *dif-1* gene to the SL splice acceptor site of *aph-1* (Fig. 1B). Such *cis*-splicing events have also been detected in putative polycistrionic RNAs discovered in tapeworm genomes (TSAI et al. 2013). Moreover, this demonstrates that *Tsp-dif-1* and *aph-1* are transcribed as a single transcript. In *T. muris* we also detected processing intermediates
corresponding to Tmu-zgpa-1-dif-1 and Tmu-dif-1-aph-1 transcripts (Figure 1B; Table S2). The latter observation is significant, since the ICR between Tmu-dif-1 and Tmu-aph-1 is predicted to be 3033 nt long, a distance substantially longer than the length of an average ICR in C. elegans; though ICRs of similar length are also present in some C. elegans operons (Morton and Blumenthal 2011).

Identification of additional putative enoplean operons

To more systematically identify enoplean operons we adopted two approaches. Firstly, we used a set of EST sequences derived from SL trans-spliced T. spiralis mRNAs (Pettitt et al. 2008) to identify their corresponding genes via sequence similarity searches (we also identified two T. muris mRNAs via the same approach: FF145866 and CB277782). For each gene we then looked for neighbouring genes, predicted to be transcribed in the same orientation and which lay within 1 kb. Using this approach we were able to identify multiple potential operons in the T. spiralis genome (Table 1; Table S1). We further analysed this set of genes by identifying the C. elegans orthologues of each T. spiralis gene and determining whether these correspond to genes within operons. Our analysis revealed that at least 75 of the C. elegans orthologues are arranged in operons. This represents 44% of the C. elegans genes identified as orthologues of our T. spiralis SL trans-spliced EST set. Since only 15% of C. elegans genes are organised into operons (Allen et al. 2011), we would expect that only 15% of T. spiralis genes in our dataset to match C. elegans operonic homologues if they were selected at random. It is difficult to determine the reason for the increased likelihood of matches to C. elegans operonic genes among the T. spiralis SL trans-spliced ESTs; it may be that the corresponding T. spiralis gene set is biased for highly expressed genes, for instance. However, it is consistent with the possibility that this dataset is enriched for transcripts derived from operonic genes.
As an alternative approach which would potentially identify operons that have been conserved since the separation of the Enoplea and Chromadoria, we looked for *T. spiralis* homologues of a set of putative operons conserved between *C. elegans* and *B. malayi* (Ghedin et al. 2007). Of the 107 operonic gene pairs screened, we identified 12 *T. spiralis* gene pairs that displayed conserved synteny, and whose component genes were separated by an average intergenic distance of 607 bp (Table 2). To determine whether any of these operons are also conserved in another enoplean species, we examined the organisation of the corresponding homologous genes in the genome of *R. culicivorax* (Schiffer et al. 2013). This analysis revealed that four of the twelve gene pairs were arranged in putative operons (assuming a maximum ICR distance of up to 1 kb) in this nematode (Table 2).

Taken together our analysis indicates that there are multiple *T. spiralis* gene pairs whose genomic arrangement is consistent with them corresponding to operons. Moreover, it is possible to identify gene pairs conserved between *T. spiralis*, *R. culicivorax*, *B. malayi* and *C. elegans*, suggesting that these represent operons that were present in the last common ancestor of the three species.

**Characterisation of a Conserved Nematode Operon**

The analysis of one of the two SL trans-spliced *T. muris* mRNAs (Genbank accession number FF145866) led to the identification of a putative operon conserved between multiple nematodes (Fig. 3A). The two genes contained in these putative operons have *C. elegans* homologues, *cpt-2* and *nuaf-3*, respectively, which are in the same operon (CEOP4424) (Fig. 3A), although there is an additional gene, *prx-14*, located between these genes in CEOP4424 that is not present in the putative homologous *T. spiralis* operon. Examination of the genomic organisation of the homologous genes in a selection of nematodes species confirmed the evolutionary conservation of the synteny of the *cpt-2* and *nuaf-3* homologues (Fig. 3A).
analysis also showed that insertion of prx-14 into the cpt-2~nuaf-3 operon was a relatively recent event, since it is present only in C. elegans and other closely related Caenorhabditis species. We also find, based on the head-to-tail organisation and spacing between coding regions that there is variation in the composition of both operons in different species, and these genes in R. culicivorax are not in operons.

We further focussed on the cpt-2~nuaf-3 operon in T. spiralis and T. muris to determine the pattern of SL trans-splicing exhibited by the mRNAs derived from this operon and to verify that we were able to detect cDNAs consistent with the production of polycistronic RNAs.

Determination of the 5’ ends of cpt-2 and nuaf-3 transcripts by 5’ RACE revealed that nuaf-3 mRNA is subject to SL trans-splicing (Table S2) in both nematodes. However, we failed to detect any SL trans-spliced cpt-2 transcripts, similar to the situation with zgpa-1 transcripts in the zgpa-1~dif-1~aph-1 operon.

We were also able to detect processing intermediates derived from the putative polycistronic T. spiralis cpt-2~nuaf-3 transcripts. As for the zgpa-1~dif-1~aph-1 operon, in addition to unprocessed, polycistronic transcripts, we detected cis-spliced intermediates lacking cpt-2 and nuaf-3 introns and a transcript in which the ICR was removed by splicing from a cryptic 3’ splice site in cpt-2 to the nuaf-3 SL splice acceptor site (Fig. 3B).

The T. spiralis cpt-2~nuaf-3 ICR can mediate polycistronic RNA processing in C. elegans

Analysis of the intercistronic region between Tsp-cpt-2 and nuaf-3 downstream of the polyadenylation signal of Tsp-cpt-2 revealed a clear Ur element and there are several U-rich regions, characteristics of the ICRs in C. elegans operons (Graber et al. 2007). To investigate the possibility that this region is able to function in polycistronic RNA processing, we determined whether the ICR from it could be recognised and processed if heterologously
expressed in *C. elegans*. We generated an artificial operon consisting of sur-5::gfp (Gu et al. 1998) and mCherry genes flanking the ICR from *Tsp-cpt-2-nuaf-3*. Transgenic animals carrying this construct expressed nuclear GFP and cytoplasmic mCherry, consistent with the processing of the two coding regions under the direction of the *Tsp-cpt-2-nuaf-3* ICR. We confirmed that this involved *trans*-splicing to SL2, as expected for polycistronic RNA processing in *C. elegans*, by showing that we could detect SL2 *trans*-splicing to the mCherry mRNA in RNA derived from transgenic animals (Fig. 4). Thus, the predicted ICR between *Tsp-cpt-2* and *nuaf-3* is recognised and used as a substrate for polycistronic RNA processing in *C. elegans*.

**Conservation of SL *trans*-splicing snRNPs between Enoplea and Chromadorea**

Our data show that both SL *trans*-splicing and operons were likely present in the last common ancestor of the nematode phylum. This suggests that the processing machinery necessary for the coordination of these processes was already in place prior to radiation of the nematode phylum. To address this, we sought to determine the conservation of known protein components that are specifically involved in SL *trans*-splicing. Previous studies have shown the existence of two interacting proteins, conserved between *A. suum* and *C. elegans*, that are components of the SL snRNP (Denker et al. 2002; MacMorriss et al. 2007). The two proteins, termed SNA-1 and SNA-2 in *C. elegans*, form a complex with SL RNA. In addition, a parologue of SNA-1, SUT-1, forms novel snRNPs containing SNA-2 and a family of nematode-specific RNAs, designated Sm Y (MacMorriss et al. 2007; Jones et al. 2009). The function of Sm Y RNAs is not known, but they are associated with SL *trans*-splicing (Maroney et al. 1996), and a role in the recycling of Sm proteins following SL *trans*-splicing has been proposed (MacMorriss et al. 2007).
We were able to identify credible homologues of sna-2 in the genomes of all nematodes in which searches were carried out, including T. spiralis and R. culicivorax, indicating that this gene encodes a nematode-wide SL trans-splicing component (Fig. S2). Searches of the same datasets identified clear sna-1 and sut-1 orthologues in the chromadorean nematodes (Fig. 5), extending the previously reported phylogenetic distribution of sut-1 (MacMorris et al. 2007). In contrast, the only enoplean genome in which we were able to identify a sut-1 homologue was R. culicivorax (Fig. 5). We could not detect sna-1 homologues in any of the enoplean genomes that we assayed. These data suggest that the gene duplication event that gave rise to sna-1 and sut-1 occurred after the separation of the two major nematode taxa, however we cannot rule out that the possibility that our failure to detect sna-1 homologues is due to the incomplete state of the enoplean genome drafts.
DISCUSSION

The incidence of operons as a means to coordinate gene expression has been investigated in only one of the two main nematode taxa, leaving unanswered the question about when this mechanism first occurred during nematode evolution (GUILIANO and BLAXTER 2006; GHEDIN et al. 2007; LIU et al. 2010). The work presented here provides strong support for the existence of operons in the Enoplea, and that operons were likely present in the ancestor of the Chromadoria and the Enoplea. The unequivocal identification of operons is not straightforward in those nematodes which do not utilise a specialised spliced leader to resolve polycistronic RNAs; the use of SL2 in C. elegans and its close relatives has greatly facilitated the identification of operons in these species. In contrast, in enoplean species we can only infer the presence of operons through multiple lines of evidence.

Our work demonstrates the existence of clusters of genes ordered in a head-to-tail arrangement with short, less than 1 kb, intergenic distances, consistent with them being organised into operons. The fact that we can identify homologous pairs of closely spaced genes conserved between enoplean and chromadorean nematodes suggests that at least in these cases there has been selective pressure to retain short intergenic distances, consistent with what would be expected if they were part of operons. Analysis of the transcripts produced by these putative operons shows that, as expected, genes predicted to be downstream in operons produce mRNAs that are SL trans-spliced. Further supporting evidence comes from the fact that we can detect unprocessed, polycistronic pre-mRNAs derived from these putative operons. Thus these genes exhibit the molecular properties expected of operons. Most significantly, the intergenic region between Tsp-cpt-2 and nuaf-3 acts as a substrate for the polycistronic RNA processing machinery of C. elegans, providing the strongest evidence that it is part of an operon in T. spiralis.
The identification of SLs from *T. muris* has extended our understanding of nematode SL evolution. Previous studies have shown that *T. spiralis* and *T. pseudospiralis* have unusually diverse SLs that do not readily correspond to those SLs found within Chromadorea (Pettitt *et al.* 2008; 2010). Another enoplean, *P. punctatus*, possesses SLs that resemble the specialised SL2 associated with *trans*-splicing and polycistronic RNA resolution found in *C. elegans* and other closely related nematodes. At least one of these is specifically recognised by the SL2 *trans*-splicing machinery in *C. elegans* when expressed heterogeneously (Harrison *et al.* 2010). Rather surprisingly, since it is more closely related to *Trichinella* species than *P. punctatus*, the *T. muris* SLs are more similar in terms of sequence composition to those of the latter nematode (the most extreme example of this being *Ppu*-SL3 and *Tmu*-SL13, which possess greater than 95% sequence identity). Thus, the diverse SL complement of the *Trichinella* species is likely to be a derived rather than ancestral trait.

In addition to elucidating the extent of nematode SL RNA conservation we have determined the phylogenetic distribution of the protein components of the SL snRNP. This reveals that one of these components, SNA-2, is conserved throughout the phylum, consistent with its essential role in SL *trans*-splicing in both *A. suum* and *C. elegans* (Denker *et al.* 2002; MacMorris *et al.* 2007). Our analysis of SNA-1/SUT-1 has revealed that these two paralogues are distributed throughout the Chromadorea. These data, together with the presence of a single clear SUT-1 homologue in *R. culicivora* (and absence of a SNA-1 homologue) is consistent with an evolutionary scenario whereby the ancestral *sut-1* gene was duplicated in the ancestor of the Chromadorea, with one paralogoue, *sut-1*, retaining the ancestral function, while the other, *sna-1*, evolved a derived function. Though, the fact that *sna-1* and *sut-1* show a synthetic loss of function phenotype in *C. elegans* (MacMorris *et al.* 2007) indicates that their functions are related and possibly overlapping. The puzzling absence of SUT-1 homologues in
*T. spiralis* and *T. muris* means that such an interpretation must be provisional pending broader sampling of genome sequences from other enopleans. If these nematodes actually lack these proteins this would be surprising given that loss of both *sna-1* and *sut-1* function in *C. elegans* results in embryonic lethality (MacMorris *et al.* 2007) and SNA-1 is required for SL \textit{trans}-splicing in *A. suum* splicing extracts (Denker *et al.* 2002).

An important question that remains to be answered is whether there is any evidence for functional specialisation of the enoplean SLs with regard to polycistronic RNA resolution. In *C. elegans*, SL2-type SLs are clearly specific for this latter process (Blumenthal *et al.* 2002), with SL1 \textit{trans}-splicing almost exclusively employed in mRNAs derived from non-operonic genes, or the first gene in an operon. Other nematodes within the Rhabditina also show the same specialisation (Evans *et al.* 1997; Lee and Sommer 2003), but other taxa within the Chromadorea appear to use the same set of SLs for all SL \textit{trans}-spliced mRNAs (Guilliano and Blaxter 2006). Within the Enoplea we clearly see multiple different SLs used in the \textit{trans}-splicing of mRNAs derived from downstream genes in operons, but our data is not comprehensive enough to determine whether some SLs are preferentially used to process the transcripts arising from such genes.

We have identified multiple examples of gene pair synteny conserved between *T. spiralis*, *R. culicivorax*, *B. malayi* and *C. elegans*, suggesting these correspond to conserved operons present in the last common ancestor of the Enoplea and Chromadorea. Nevertheless, the majority of the putative operons that we have identified in *T. spiralis* are not syntenic with *C. elegans* operons, though in many cases we see that the component genes of these putative *T. spiralis* operons have *C. elegans* orthologues that are found in operons. An example of such an operon is that comprising *Tsp-zgpa-1*, *Tsp-dif-1* and *Tsp-aph-1*, whose *C. elegans* orthologues are located in different operons. We also observed putative *T. spiralis* operon genes whose *C.
*elegans* orthologues are not part of operons. The changes in the operon complements between the two nematodes could arise from the lineage-specific rearrangements of an ancestral set of nematode operons, but could also be accounted for by differential *de novo* operon generation in the two clades, or a mixture of the two processes. A key question is whether the synteny shown between the putative enoplean operons and their chromadorean orthologues is significant, i.e. are the genes that comprise these operons more constrained to be located in the same operon than other operonic genes, or is the conservation of synteny merely random chance? The availability of tools to engineer the *C. elegans* genome (Golic 2013) might allow this question to be addressed by assaying the function of selected operons compared with their individual component genes each expressed under their own promoters.

Finally, it is clear from this work and previous studies that SL *trans*-splicing and operon organisation arose prior to the divergence of the Enoplea and Chromadorea. An important question is whether "nematode" SL *trans*-splicing and operons predate the foundation of the phylum. It will thus be important to establish whether they are also present in the other, so far uncharacterised, phyla that are closely related to the nematodes.
ACKNOWLEDGMENTS

Thanks to Allison Bancroft and Richard Grencis for providing *T. muris* RNA. The Bristol (N2) *C. elegans* N2 strain was provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). This work was funded by a BBSRC project grant BB/J007137/1.

FOOTNOTES

The *T. spiralis* and *T. muris* mRNA 5’ ends with/without SL sequences are deposited in GenBank (accession numbers KF442418-KF442435; KF511776; KF511777; KF768019).
REFERENCES


JONES T., OTTO W., MARZ M., EDDY S., STADLER P., 2009  A survey of nematode SmY RNAs. RNA Biol **6**.


26


Table 1. Operonic Status of Genes that match ESTs derived from *T. spiralis* SL trans-spliced transcripts

<table>
<thead>
<tr>
<th>Species</th>
<th>Location in operon</th>
<th>Non-operonic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Upstream</td>
<td>Downstream</td>
</tr>
<tr>
<td><em>T. spiralis</em></td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>29</td>
<td>46</td>
</tr>
</tbody>
</table>

The status of each *T. spiralis* gene was determined using criteria given in Materials and Methods. The status of *C. elegans* genes was obtained from Wormbase (Release WS237). Eleven EST matches were absent from the *T. spiralis* data relative to the *C. elegans* data as the corresponding *T. spiralis* gene could not be identified (see Materials and Methods). Full details of the individual EST sequence matches are given in Table S1.

* Indicates genes not annotated as operons, but having intergenic spacing with respect to their neighbours that suggests they may be organised in an operon.
Table 2. Putative conserved nematode operons

<table>
<thead>
<tr>
<th><strong>T. spiralis</strong></th>
<th><strong>R. culicivorax</strong></th>
<th><strong>C. elegans</strong></th>
<th><strong>B. malayi</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Upstream Gene</td>
<td>Downstream Gene</td>
<td>Upstream Gene</td>
<td>Downstream Gene</td>
</tr>
<tr>
<td>Tsp_00685</td>
<td>*</td>
<td>Non operonic</td>
<td>mrps-17</td>
</tr>
<tr>
<td>Tsp_03140</td>
<td>Tsp_03139</td>
<td>t32947</td>
<td>T26E3.4</td>
</tr>
<tr>
<td>Tsp_05540</td>
<td>Tsp_05541</td>
<td>Non operonic</td>
<td>K11B4.1</td>
</tr>
<tr>
<td>Tsp_06077</td>
<td>Tsp_06076</td>
<td>Non operonic</td>
<td>Y62E10A.2</td>
</tr>
<tr>
<td>Tsp_06996</td>
<td>*</td>
<td>t05598/9</td>
<td>sel-1</td>
</tr>
<tr>
<td>Tsp_09103</td>
<td>Tsp_09102</td>
<td>t35569</td>
<td>snu-23</td>
</tr>
<tr>
<td>Tsp_09506</td>
<td>*</td>
<td>Non operonic</td>
<td>H20J04.6</td>
</tr>
<tr>
<td>Tsp_09539</td>
<td>*</td>
<td>Non operonic</td>
<td>E02H1.5</td>
</tr>
<tr>
<td>Tsp_10673</td>
<td>Tsp_10674</td>
<td>t34344.1</td>
<td>B0491.1</td>
</tr>
<tr>
<td>Tsp_10698</td>
<td>Tsp_10702</td>
<td>Non operonic</td>
<td>trpp-8</td>
</tr>
<tr>
<td>Tsp_10959</td>
<td>*</td>
<td>Non operonic</td>
<td>ubxn-2</td>
</tr>
<tr>
<td>Tsp_11898</td>
<td>*</td>
<td>Non operonic</td>
<td>ist-6</td>
</tr>
</tbody>
</table>

* Single gene annotation matches to both *C. elegans* genes in the operon pair consistent with annotation error caused by short intergenic spacing.
Figure 1. Evidence for the existence of an enoplean operon. A. Schematic showing the genomic organisation of *zgpa-1*, *dif-1* and *aph-1* in selected nematodes mapped onto their phylogenetic relationships. Arrows represent genes, and the grey lines represent the intercistronic regions (ICRs). Numbers above the ICRs represent the distances, in base pairs, between the stop and start codons of the upstream and downstream genes, respectively. The *C. elegans* operon numbers are given where appropriate. Fractions below the *T. spiralis* and *T. muris* genes represent the proportion of cDNAs derived from that gene that begin with a spliced leader sequence (see also Table S2). In *C. elegans*, the three genes are part of different operons. * indicates the distance between genes on chromosome IV. (B) Detecting polycistronic RNAs derived from the *zgpa-1*-*dif-1*-*aph-1* operon in enoplean nematodes. The exon-intron structures of the amplicons used to identify polycistronic RNAs are shown, with exons represented by boxes (shaded to identify the genes from which they are derived using the same colour coding that was used in A). The intercistronic regions are represented by cream coloured boxes. The positions of the SL trans-spliced 3’ splice sites are indicated. The length of each cDNA is indicated.
Figure 2. *T. muris* SL sequences and SL RNA structure. (A) *Tmu*-SL1 to *Tmu*-SL13 genes were identified using a combination of cDNA sequencing and bioinformatics tools as described in Materials and Methods. *Tmu*-SL12 was found by 5’ RACE trans-spliced to *nuaf-3* mRNA, and *Tmu*-SL13 was found trans-spliced to *aph-1* mRNA. In the alignment only the SL sequences are shown. *T. muris* SL sequences were manually aligned and conserved groups are counter-shaded. *C. elegans* SL1 and SL2 and the previously identified *P. punctatus* SL sequences were included for comparison. (B) The intron of *Tmu*-SL2 was experimentally identified, and also found in the genome sequence. The proposed secondary structure was produced using M-fold (Zuker 2003). The SL sequence is shown in bold and the putative Sm sequence motif is counter-shaded.
Figure 3. Evidence for an evolutionarily conserved nematode operon. (A) The structure of the cpt-2-nuaf-3 genomic regions from a range of nematode species mapped onto the nematode phylogeny. Genes are represented by arrows, and the grey lines represent the intercistronic regions (ICRs). Numbers above the ICRs represent the distances, in base pairs, between the stop and start codons of the upstream and downstream genes, respectively. The C. elegans operon numbers are given where appropriate. Fractions below the T. spiralis and T. muris genes represent the proportion of cDNAs derived from that gene that begin with a spliced leader sequence. (B) Detecting polycistronic RNAs derived from the cpt-2~nuaf-3 operon in T. spiralis. The exon-intron structures of the amplicons used to identify polycistronic RNAs are shown, with exons represented by boxes (shaded to identify the genes from which they are derived using the same colour coding that was used in A). The region removed during operon processing is represented by cream coloured boxes. The positions of the SL trans-splice 3’ spliced sites are indicated.
Figure 4. Processing of a synthetic operon containing the T. spiralis cpt-2-nuaf-3 intercistronic region in C. elegans. (A) Schematic of the structure of sur-5::gfp~mCherry synthetic operon construct containing the Tsp-cpt-2-nuaf-3 intercistronic region. The sequence immediately downstream from the Tsp-cpt-2 3'UTR is shown, illustrating the presence of the Ur motif, and the trans-splice acceptor site. Strain PE613 contains an identical construct, but with the ICR replaced with that from between Cel-cpt-2 and prx-14. (B) Detection of operon transcript processing by SL2 trans-splicing. mCherry transcripts trans-spliced to SL2 were detected by reverse transcription of RNA prepared from either PE612 or PE613 animals followed by PCR with an SL2 primer and a primer located in the mCherry coding region (+RT). Primers amplifying gpd-1 were included to control for sample variation. gpd-1 genomic DNA was detected in the -RT control reaction (*) and SL2-ZK1236.7a is a minor product detected in the +RT reactions. Reactions with RNA isolated from N2 wild type animals were included as control. M is a DNA size standard. (C) Alignment of transgene sequences and SL2-mCherry transcripts confirming correct splice site usage. The beginning of the mCherry open reading frame is counter shaded black, the Nhel cloning site is counter shaded grey, and the SL2 sequences are in bold.
Figure 5. Evolutionary relationship between snRNPs associated with SL trans-splicing.
Unrooted PhyML tree showing the relationship between sna-1 and sut-1 homologues identified in selected nematodes. Genes were named on the basis of their C. elegans homologues. The following species-specific prefixes were used: Aca – Angiostrongylus cantonensis; Asu – Ascaris suum; Bma – Brugia malayi; Cel – C. elegans; Cbr – C. briggsae; Cre – C. remanei; Hco – Haemochus contortus; Llo – Loa loa; Nam – Necator americanus; Rcu – Romanomermis culicivorax; Wba – Wuchereria bancrofti. The numbers at each node are aLRT statistics.