Resolution of polycistronic RNA by SL2 *trans*-splicing is a widely conserved nematode trait

MARIUS WENZEL,[1,3] CHRISTOPHER JOHNSTON,[2] BERNDT MÜLLER,[2] JONATHAN PETTITT,[2] and BERNADETTE CONNOLLY[2]

[1]Centre of Genome-Enabled Biology and Medicine, University of Aberdeen, Aberdeen AB24 3RY, United Kingdom

[2]School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Institute of Medical Sciences, Foresterhill,
Aberdeen AB25 2ZD, United Kingdom

ABSTRACT

Spliced leader *trans*-splicing is essential for the processing and translation of polycistronic RNAs generated by eukaryotic operons. In *C. elegans*, a specialized spliced leader, SL2, provides the 5′ end for uncapped pre-mRNAs derived from polycistronic RNAs. Studies of other nematodes suggested that SL2-type *trans*-splicing is a relatively recent innovation, confined to Rhabditina, the clade containing *C. elegans* and its close relatives. Here we conduct a survey of transcriptome-wide spliced leader *trans*-splicing in *Trichinella spiralis*, a distant relative of *C. elegans* with a particularly diverse repertoire of 15 spliced leaders. By systematically comparing the genomic context of *trans*-splicing events for each spliced leader, we identified a subset of *T. spiralis* spliced leaders that are specifically used to process polycistronic RNAs—the first examples of SL2-type spliced leaders outside of Rhabditina. These *T. spiralis* spliced leader RNAs possess a perfectly conserved stem–loop motif previously shown to be essential for SL2-type *trans*-splicing in *C. elegans*. We show that genes *trans*-spliced to these SL2-type spliced leaders are organized in operonic fashion, with short intercistronic distances. A subset of *T. spiralis* operons show conservation of synteny with *C. elegans* operons. Our work substantially revises our understanding of nematode spliced leader *trans*-splicing, showing that SL2 *trans*-splicing is a major mechanism for nematode polycistronic RNA processing, which may have evolved prior to the radiation of the Nematoda. This work has important implications for the improvement of genome annotation pipelines in nematodes and other eukaryotes with operonic gene organization.

Keywords: **operons; *trans*-splicing; nematodes; *Trichinella spiralis*; polycistronic RNA**

INTRODUCTION

The organization of multiple genes into a single transcriptional control unit, termed an operon, is a common, though sparsely distributed, gene expression strategy in eukaryotes (Blumenthal 2004; Douris et al. 2010; Danks et al. 2015). Although the general adaptive significance of this mode of gene organization remains uncertain (Zaslaver et al. 2011; Danks et al. 2015), a common feature in all cases is the presence of spliced leader *trans*-splicing. This derived version of *cis*-splicing allows the addition of short, "spliced leader" RNAs to the 5′ ends of pre-mRNAs via an intermolecular splicing event (Lasda and Blumenthal 2011). It is a critical element in the generation of mRNAs derived from genes situated downstream from the first gene in an operon, since the spliced leader RNA provides the 5′ cap for such mRNAs, allowing the *trans*-spliced mRNAs to be recognized by the translation machinery. Addition of the spliced leader is also thought to prevent termination of transcription following polyadenylation of the upstream mRNA (Evans et al. 2001; Haenni et al. 2009; Lasda et al. 2010). Thus, spliced leader *trans*-splicing significantly facilitates the evolution of eukaryotic operons.

Spliced leader *trans*-splicing and operon organization are best understood in the nematode *C. elegans* (Blumenthal 2012; Blumenthal et al. 2015). At least 84% of *C. elegans* genes encode mRNAs that are spliced leader *trans*-spliced, with ∼9% of these arising from downstream genes in operons (Allen et al. 2011; Tourasse et al. 2017). Two functionally distinct types of spliced leaders are

[3]**Present address:** School of Biological Sciences, University of Aberdeen, Aberdeen AB24 2TZ, United Kingdom

Corresponding author: j.pettitt@abdn.ac.uk

Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.076414.120.

present in *C. elegans*. The first to be discovered, SL1, is *trans*-spliced to pre-mRNAs derived from monocistronic genes and the first genes in operons (Krause and Hirsh 1987). SL1 *trans*-splicing likely serves to sanitize the 5′ ends of mRNAs, removing deleterious sequences from 5′ untranslated regions, and thus impacts mRNA translational efficiency (Yang et al. 2017). The other *C. elegans* spliced leader type, SL2, is added to mRNAs encoded by downstream operonic genes, which are otherwise uncapped and thus cannot be translated without SL2 *trans*-splicing (Huang and Hirsh 1989; Spieth et al. 1993). Multiple SL2 isoforms have been identified and these are encoded by 18 genes (MacMorris et al. 2007).

The two spliced leader RNA types both fold into three stem–loop structures but have differentiated biochemical interactions (Evans et al. 2001; MacMorris et al. 2007). SL2 RNAs have a motif present in the third stem–loop that is not present in SL1 RNA (Evans and Blumenthal 2000). The sequence composition of this stem–loop is essential for the specificity of SL2 *trans*-splicing to downstream operonic gene pre-mRNAs, as well as its association with Cleavage Stimulation Factor Subunit 2 (CSTF2), a principal factor involved in coordinating polyadenylation of the upstream pre-mRNA with the spliced leader *trans*-splicing of the downstream pre-mRNA (Evans et al. 2001).

While SL2-type *trans*-splicing is found in other species belonging to the same subclass as *C. elegans* (Rhabditina or Clade V) (Evans et al. 1997; Lee and Sommer 2003), previous studies failed to detect SL2-type *trans*-splicing in the other major nematode clades (Guiliano and Blaxter 2006; Ghedin et al. 2007). This led to the view that SL2 *trans*-splicing is a relatively recent innovation, associated with more efficient processing of polycistronic RNAs (Blumenthal 2012) that evolved only in one nematode lineage. However, the limited exploration of spliced leader RNA usage in the other major nematode clades means that this hypothesis has yet to be rigorously investigated.

We have previously studied spliced leader *trans*-splicing and operons in *Trichinella spiralis* (Pettitt et al. 2008, 2014), a nematode which belongs to a clade, Dorylaimia (Clade I), that diverged from the lineage leading to *C. elegans* early on during the radiation of the Nematoda. The genome of *T. spiralis* encodes at least 15 distinct spliced leader RNAs, whose nucleotide sequences show a high degree of polymorphism, and no sequence similarity with *C. elegans* SL1 or SL2s (Pettitt et al. 2008). The sequence diversity of these SL RNAs might simply be a consequence of unconstrained sequence variation but could also reflect functional differences.

Here we describe the results of a transcriptome-wide investigation of spliced leader usage in the muscle larva of *T. spiralis*. We show that this nematode possesses a subset of spliced leaders that, like *C. elegans* SL2s, are specialized for the processing of pre-mRNAs derived from downstream genes in operons. These spliced leaders share a motif in their third stem–loops that is identical to those found in *C. elegans* SL2s, suggesting that their specificity arises from the same conserved interaction with the polyadenylation machinery. This feature is found in spliced leader RNAs in members of multiple nematode lineages. Thus, rather than being a recent innovation found only in a subset of nematodes, SL2-type spliced leader *trans*-splicing is broadly distributed throughout the phylum.

RESULTS

A minority of *T. spiralis* genes are subject to spliced-leader *trans*-splicing

In order to explore the possibility that sequence diversity among the 15 *T. spiralis* spliced leaders reflects possible functional diversification, we investigated the patterns of spliced leader usage in the muscle larva transcriptome. We sequenced the transcriptomes of three independent pools of *T. spiralis* L1 muscle larvae using Illumina RNA-seq and identified reads containing spliced leaders (hereafter: *Tsp*-SLs) at their 5′ end. We generated 201,164,867 high-quality read pairs across all pools (Supplemental Table S1, sheet “Read statistics”). Of these read pairs, 84.2% aligned concordantly end-to-end to the *T. spiralis* reference genome, whereas 2.9% contained a single read that did not and could thus contain a *trans*-spliced leader sequence. Of these unaligned reads, 6.1% (0.2% of all read pairs) were unambiguously assigned to one of 15 known *Tsp*-SL types, matching at least 10 bp at the 3′ end of the characteristic spliced leader sequences (Supplemental Table S1, sheet “Read statistics”). The numbers of read pairs per *Tsp*-SL type varied considerably, ranging from 1290 (*Tsp*-SL9) to 58,838 (*Tsp*-SL11) (Supplemental Table S1, sheet “Read statistics”). An alternative *Tsp*-SL read screening pass, using a more relaxed 8 bp minimum overlap could no longer reliably distinguish between *Tsp*-SL13, 14, and 15, but increased the number of reads identified to 8.9% of candidate read pairs (0.3% total read pairs) and the number of read pairs per *Tsp*-SL type from 1722 (*Tsp*-SL9) to 97,730 (*Tsp*-SL6) (Supplemental Table S1, sheet “Read statistics”).

Our ability to accurately identify the genes that receive each *Tsp*-SL type depends critically on reliable gene annotations. Since the draft genome annotations for *T. spiralis* are not based on transcriptomic evidence (Mitreva et al. 2011), we decided to reannotate the genome de novo using the concordant read-pair alignments generated from our three RNA-seq libraries. This yielded 13,060-20,083 genes across four different annotation pipelines (see Materials and Methods). The *T. spiralis* reference annotation contains 16,380 genes composed of 87,853 nonredundant exons; our RNA-seq reads covered 15,025 of these genes, suggesting that 92% of the reference genes are expressed in *T. spiralis* L1 muscle larvae. However, no more than 49.6% of reference exons and 10.8% of reference transcripts matched our de novo annotation sets, indicating that the reference annotation does not accurately reflect the gene structures suggested by our RNA-seq data. This was further confirmed using BUSCO, where completeness of the reference annotations was lower (67.0%) than any of the four de novo annotation sets (68.0%–71.5%; Supplemental Table S2). Of these, the BRAKER + TRINITY pipeline (see Materials and Methods) yielded the lowest number of duplicated orthologs and most accurate gene models when inspecting known operonic genes (see details below).

The identified *Tsp*-SL read pairs aligned back to the genome at rates of 91.6%–98.9% (properly paired 79.2%–97.0%), and the *Tsp*-SL-containing single-end read alignments were unambiguously assigned to single exon annotations at rates of 48.8%–73.5% for the reference annotations, and 53.8–88.0 for de novo annotations (Supplemental Table S1, sheet "Alignment"). Given correct gene annotations, *Tsp*-SL reads are expected to align to the first exon of the *trans*-spliced gene. This was indeed the case for most *Tsp*-SL reads, but in a number of cases internal exons received distinct peaks of *Tsp*-SL reads, indicating that single gene annotations were actually composites of two or more genes. Adjusting gene numbers based on this evidence led to between 250 to 1103 extra genes per data set (Supplemental Table S2), and manual inspection of known operonic genes confirmed that these adjustments were accurate. Since this suggests considerable issues regarding the accuracy of gene annotations even in our de novo annotation sets, we carried out all further analyses on 20 data sets in total (8/10 bp minimum tail overlap during *Tsp*-SL screening, five gene annotation sets, and using either uncorrected or *Tsp*-SL-corrected gene annotations). This allowed us to examine the sensitivity of our results to these three variables throughout all analyses. Across all these 20 data sets, the percentage of *T. spiralis* genes whose mRNAs are spliced leader *trans*-spliced ranged from 18.7% to 31.3% (Supplemental Table S2). These observations are substantially lower than the 80%–90% of genes reported for other nematode species (Maroney et al. 1995; Allen et al. 2011; Sinha et al. 2014).

Trichinella spiralis SL RNAs are structurally and functionally differentiated

In silico examination of the secondary structure of the 15 *Tsp*-SL RNAs revealed a distinct motif at the top of the third stem–loop in *Tsp*-SL2, SL10, and SL12 RNAs, which was absent from the other 12 RNAs (Fig. 1A). This motif is identical to that found in *C. elegans* SL2 RNAs, which is essential for the specialized activity of this family of spliced leaders (Evans and Blumenthal 2000; Evans et al. 2001). Inspection of the predicted structures of spliced leader RNAs from other Clade I nematodes, *Trichinella pseudospiralis*, *Trichuris muris*, *Soboliphyme baturini*, and *Prionchulus punctatus* as well as Clade IV nematodes, *Bursaphelenchus xylophilus* and *Panagrellus redivivus*, showed that this motif is broadly conserved in the predicted third stem–loops of a subset of nematode spliced leader RNAs (Fig. 1A; Supplemental Fig. S1). The striking structural conservation of this motif suggests that *Tsp*-SL2, SL10, and SL12 spliced leader RNAs share the functional properties of *C. elegans* SL2 RNAs, and we predict that they would be associated with mRNAs derived from a distinct set of genes (downstream operonic genes) compared to the other *Tsp*-SLs.

To investigate this in an unbiased way, we determined whether each of the annotated *T. spiralis* genes (13,060–20,676 depending on annotation set) was represented (presence/absence) by each individual *Tsp*-SL read set (*Tsp*-SL1 through *Tsp*-SL15). In this way, patterns of genome-wide alignment locations of each *Tsp*-SL read set are represented as sequences of binary variables corresponding to each gene. Multivariate analysis of these data, using Jaccard distances to measure the dissimilarity between the read sets, shows that the 15 spliced leader read sets could be grouped into two major clusters, one composed of *Tsp*-SL2, SL10, and SL12, and the other containing all other spliced leaders (Fig. 1B,C). Alignment patterns between the three replicate sequencing libraries were highly consistent for all *Tsp*-SL read sets (Fig. 1B,C). To examine the number of genes that contribute to the distinction between the two spliced leader read clusters, we performed linear discriminant analysis. The single discriminant function strongly separated the two clusters with no overlap, and the genes most strongly driving this separation were distributed across all genomic scaffolds (Fig. 1D). Thus, we see a strong, genome-wide difference between the patterns of *trans*-splicing displayed by *Tsp*-SL2, SL10, and SL12 compared to the other *Tsp*-SLs.

These patterns were broadly consistent across all gene annotation sets and also when accounting for read depth in dissimilarity calculations. However, the degree of similarity between *Tsp*-SL2 and *Tsp*-SL10/SL12 varied such that *Tsp*-SL2 was sometimes not clustered with *Tsp*-SL10/SL12, particularly when accounting for read depth (Supplemental Fig. S2). This suggests that although they share similar activity, *Tsp*-SL2 RNAs may have some distinct properties compared to *Tsp*-SL10 and *Tsp*-SL12 RNAs. Nevertheless, the data show that *Tsp*-SL2, SL10, and SL12 are functionally distinct from the other *T. spiralis* spliced leaders, consistent with the possibility that they are mechanistically distinct as expected for SL2-type spliced leaders.

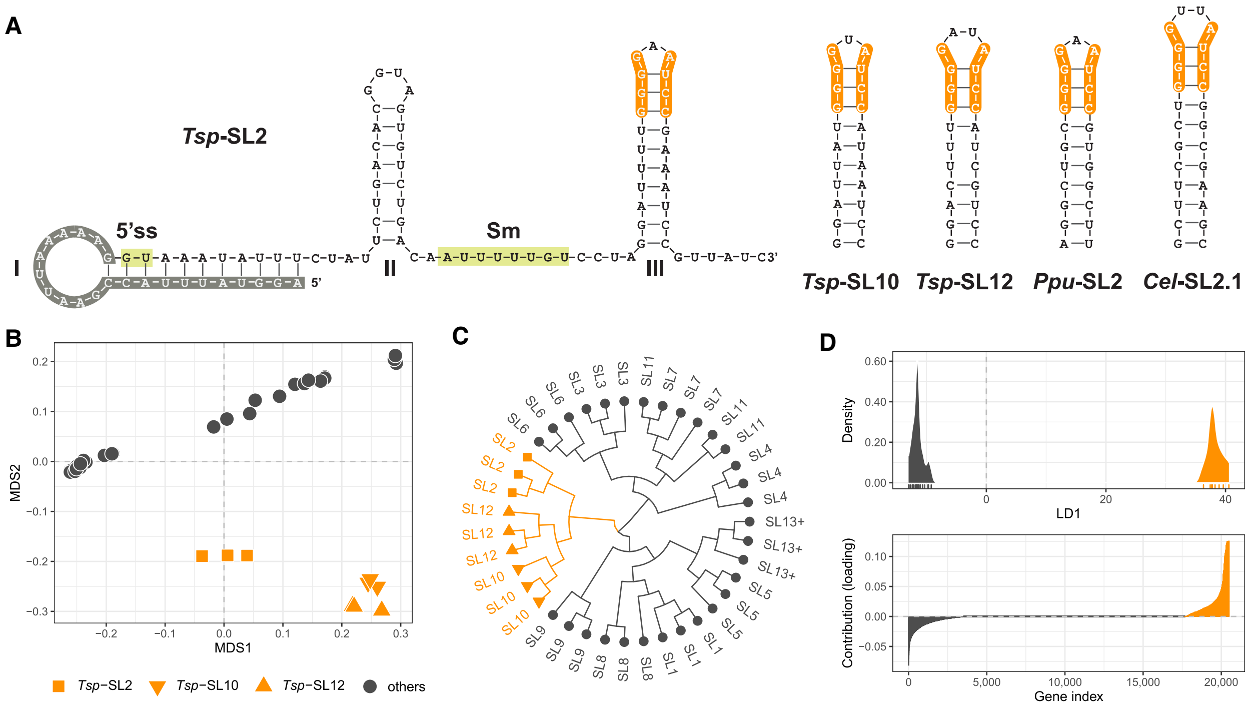

FIGURE 1. *Tsp*-SL2, *Tsp*-SL10, and *Tsp*-SL12 RNAs are structurally and functionally distinct from other *T. spiralis* spliced leader RNAs. (*A*) Secondary structure predictions for the *T. spiralis* SL (*Tsp*-SL) RNAs that possess a conserved third stem–loop motif shared with *C. elegans* SL2.1 RNA (orange shading). Full length *Tsp*-SL2 is shown to indicate the three numbered stem–loops (I, II, III), spliced leader sequence (gray), donor splice site (5′ss; green), and Sm-binding site (green). Only the third stem–loops are shown for *Tsp*-SL10 and SL12, and *P. punctatus* SL2. (*B*) Multidimensional scaling (MDS) plot of Jaccard distances among *Tsp*-SL read sets (classified with 8 bp minimum match) from three replicate RNA-seq libraries (*Tsp*-SL1 to *Tsp*-SL15) based on presence/absence of spliced leader reads at each gene in the *T. spiralis* genome (BRAKER + TRINITY *Tsp*-SL-corrected gene annotations; see main text). (*C*) Hierarchical clustering (Ward's method) of gene-based Jaccard distances among *Tsp*-SL read sets (*Tsp*-SL is abbreviated to "SL" for simplicity). Note that SL13+ contains *Tsp*-SL13, *Tsp*-SL14, and *Tsp*-SL15, which are not reliably distinguishable with an 8 bp minimum tail match. (*D*) Graphical overview of gene-specific contributions to multivariate group separation (linear discriminant analysis) between *Tsp*-SL read sets containing *Tsp*-SL2, SL10, or SL12 (orange) versus all other SLs (gray). The score of each read set in the linear discriminant function (LD1) is plotted as rug marks along the *x*-axis and score densities within the two groups are overlaid on the *y*-axis. *Below*, the contribution (variable loading) of each gene to group discrimination along LD1 is plotted and colored according to directionality (mathematical sign). Genes are ordered by contribution in ascending fashion.

mRNAs derived from putative downstream genes in *T. spiralis* operons are *trans*-spliced almost exclusively to *Tsp*-SL2, *Tsp*-SL10, and *Tsp*-SL12

Having established that *Tsp*-SL2, SL10, and SL12 are functionally distinct from the other *T. spiralis* spliced leader RNAs, we investigated whether this was because they are specifically associated with downstream genes in *T. spiralis* operons. To do this, we examined the *Tsp*-SL read coverage across upstream and downstream genes in a set of 45 manually curated, putative *T. spiralis* operons, identified on the basis of their syntenic conservation with *C. elegans* operonic gene pairs (Pettitt et al. 2014; Johnston 2018) and unusually short intergenic distances (Supplemental Table S3).

Strikingly, we observed that upstream and downstream operonic genes showed distinct distributions of *Tsp*-SLs (Fig. 2A). Transcripts from 48 of the 59 downstream genes in our set of manually curated operons were exclusively *trans*-spliced to *Tsp*-SL2, SL10, and SL12, with the remainder showing a strong bias (read ratio of *Tsp*-SL2/SL10/SL12 to all other *Tsp*-SLs ≥ 1; median of 14) for these three spliced leaders (Fig. 2B,C). Upstream genes showed an opposite, though less extreme bias in SL usage, with a preference for *Tsp*-SLs other than *Tsp*-SL2, SL10, and SL12 (Fig. 2B,C). These data provide strong evidence that *Tsp*-SL RNAs fall into two classes, functionally equivalent, if not homologous, to *C. elegans* SL1- and SL2-type SL RNAs. Since *Tsp*-SL2, SL10, and SL12 are primarily added to downstream operon transcripts, we thus refer to them as "SL2-type" spliced leaders (Fig. 2B,C; Supplemental Table S3), and all the other *Tsp*-SL RNAs as "SL1-type" spliced leaders (Fig. 2B,C; Supplemental Table S3). Since the gene annotations derived from the BRAKER + TRINITY pipeline with exon-based correction agreed best with these well-characterized operons, the following results are reported primarily for this data set, where appropriate.

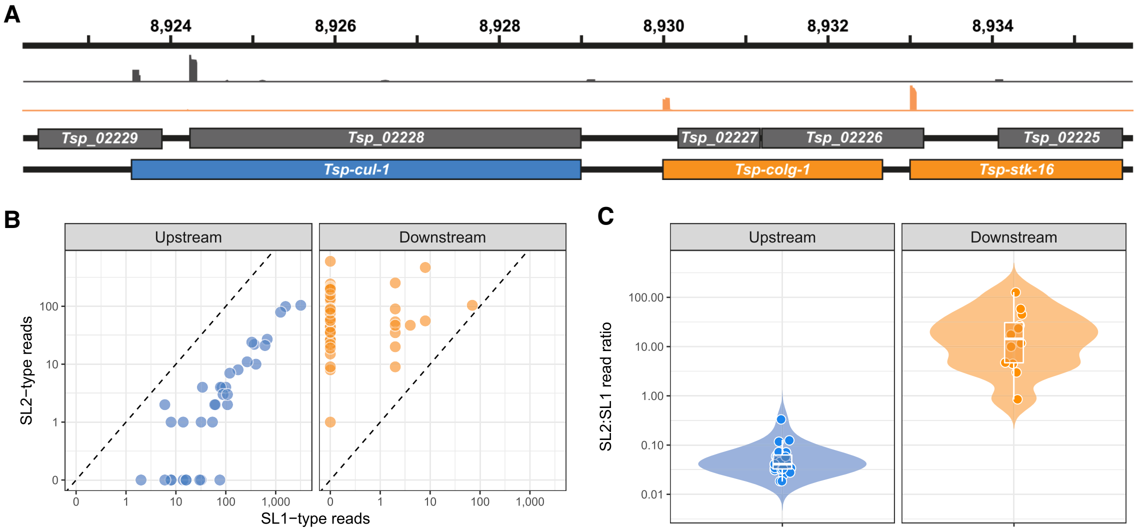

FIGURE 2. *Tsp*-SL2, *Tsp*-SL10, and *Tsp*-SL12 *trans*-splicing defines downstream genes in operons. (*A*) Example of a manually curated *T. spiralis* operon. Revision of the original gene predictions (represented as dark shaded boxes) shows this to be a three gene operon. Alignment of *Tsp*-SL-containing reads shows differential usage of spliced leaders, with *Tsp*-SL2, *Tsp*-SL10, and *Tsp*-SL12 containing reads (defined as SL2-type reads, orange peaks) being greatly enriched at the downstream genes. Conversely, all other *Tsp*-SLs (SL1-type reads, gray peaks) are enriched at upstream genes. Depth for both tracks is scaled to 80. The revised *T. spiralis* gene annotations are named based on their *C. elegans* and/or human orthologs. Note that the exon/intron structure of the genes is not shown. Many of the SL1-type reads span an intron in *Tsp-cul-1*. X-axis indicates distance in kilobases along Scaffold GL622787. (*B*) Scatter plots of SL1- and SL2-type read counts of *trans*-splice sites for genes upstream or downstream in 45 manually curated operons. Upstream genes without any *Tsp*-SL reads are not plotted. (*C*) Distributions of SL2:SL1 read-count ratios for genes upstream or downstream in operons. Each gene is represented by a dot and distributions are summarized with boxplots and density plots. Only genes with at least one SL1 and SL2 read are plotted.

Tsp-SL2, *Tsp*-SL10, and *Tsp*-SL12 *trans*-splicing enables genome-wide prediction of *T. spiralis* operons

The demonstration that mRNAs derived from downstream operonic genes in *T. spiralis* are *trans*-spliced to a specialized set of SL2-type spliced leaders meant that we could use the distributions of SL1-type and SL2-type *trans*-splicing events as a diagnostic tool to predict operons genome-wide. The majority of spliced leader *trans*-spliced genes were spliced exclusively to SL1-type spliced leaders and a minority received exclusively SL2-type spliced leaders (Fig. 3A). A third category comprised genes that received a mixture of both SL types, with similar gene numbers to the "SL1-type only" category (Wilcoxon rank-sum test: $P = 0.24$). Genes in this mixed category displayed the highest median read depth ($P < 0.001$; Fig. 3B) and the same strong bias toward SL1-type reads (median SL2:SL1 read ratio: 0.11; Fig. 3C) that we had already observed among our 45 control operons (Fig. 2). The suspiciously low read depth in the "SL1-type only" category suggests that this category may mostly be an artefact, that is, there were too few reads to detect any rare SL2-type reads (Fig. 3B). The same was not the case for the "SL2-type only" category (downstream operonic genes), where read depths were intermediate and the top 100 genes with the highest SL2-type read depth had a geometric mean of only 0.43 SL1-type reads versus 198.21 SL2-type reads. Thus, it appears that SL1-type spliced leader RNAs are very poor substrates for the *trans*-splicing of pre-mRNAs from downstream operonic genes, but monocistronic or upstream operonic genes are much more tolerant of SL2-type *trans*-splicing. Among these "SL1 + SL2" genes, *Tsp*-SL2 was much more frequently used than *Tsp*-SL10 and *Tsp*-SL12 (Fig. 3D), whereas SL2-type genes used *Tsp*-SL2 much more rarely and instead were biased toward *Tsp*-SL12 (Fig. 3D). These results echo the functional distinction between the SL2-type spliced leader RNAs (*Tsp*-SL2 in particular) observed in the multivariate exploration (Supplemental Fig. S2). It is, however, worth noting that operonic genes receiving all three SL2-type spliced leaders were the most frequent category overall (Fig. 3D).

On the basis of these observations, we designated uninterrupted runs of genes that were spliced either exclusively to or showed strong preference for SL2-type spliced leaders (SL2:SL1 > 2) as "downstream" in operons, and thus identified 477 operons comprising 992 genes using the most accurate gene annotations. Consistent with the low proportion of genes whose mRNAs were enriched for SL2-type spliced leaders (Fig. 3A,C), the numbers of inferred operonic genes were much smaller than those of *trans*-spliced nonoperonic (monocistronic) genes (Fig. 4A). Predicted operon numbers varied considerably depending on the gene annotation set used, ranging

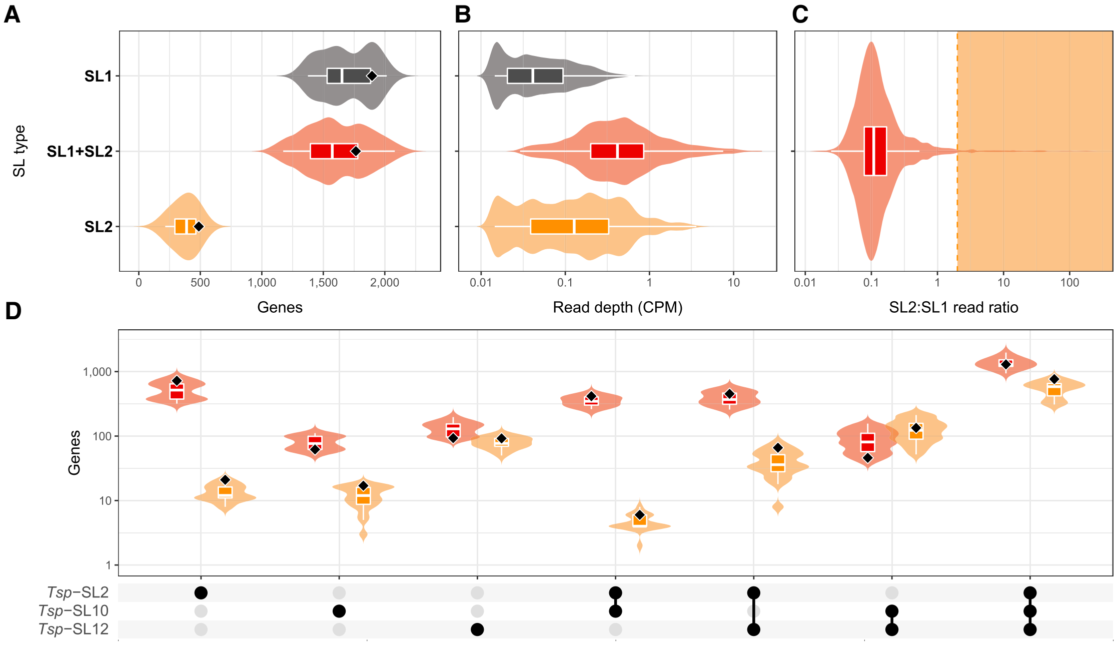

FIGURE 3. Genome-wide identification of differential usage of *T. spiralis* spliced leader RNAs. Genome-wide identification of *T. spiralis* genes receiving either exclusively SL2-type spliced leaders (*Tsp*-SL2, *Tsp*-SL10, and *Tsp*-SL12), SL1-type spliced leaders (all other *Tsp*-SLs), or a mixture of both types ("SL1 + SL2"). (*A*) Numbers of expressed genes receiving *Tsp*-SL reads. Plots show distributions (density and boxplot) across all 20 data sets in total (see main text); the black diamonds indicate the value for the most accurate data set (BRAKER + TRINITY, classified with 8 bp minimum match). (*B*) Distributions of *Tsp*-SL read depths (counts per million; CPM) using the most accurate data set (logarithmic scale). (*C*) Distribution of SL2:SL1 read-depth ratio in genes receiving both SL1- and SL2-type spliced leaders, using the most accurate data set (logarithmic scale). Genes with ratio >2 (orange shaded range of the distribution) were considered SL2-type genes for operon prediction consistent with observations from known benchmark operons (Fig. 2). (*D*) Distributions of gene numbers among "SL1 + SL2"-type (red) and "SL2"-type (orange) genes receiving combinations of *Tsp*-SL2, *Tsp*-SL10, and *Tsp*-SL12 spliced leaders, across all 20 data sets (black diamonds indicate value for the most accurate data set). The *x*-axis represents all possible intersections among the three *Tsp*-SLs as a combination dot matrix.

between 275 and 509 operons, defining between 551 and 1045 operonic genes (Supplemental Table S2). Correcting gene annotations by splitting genes at internal exons that receive distinct peaks of *Tsp*-SL reads significantly increased (Wilcoxon rank-sum test: $P < 0.005$) the numbers of *trans*-spliced genes, operons and operonic genes (Supplemental Table S2). The poor-quality reference gene annotations yielded marginally fewer ($P = 0.064$) *trans*-spliced genes than the de novo annotations, though numbers of operons and operonic genes were not significantly lower (Supplemental Table S2). Relaxing the minimum tail match during *Tsp*-SL screening from 10 bp to 8 bp yielded significantly more reads ($P < 0.001$) and *trans*-spliced genes ($P = 0.019$), but not significantly more operons or operonic genes (Supplemental Table S2).

Caenorhabditis elegans genes within the same operon have characteristically short intergenic distances compared to monocistronic genes (Allen et al. 2011). We observe a similar trend for *T. spiralis* genome-wide predicted operons, based on the distance between the stop codon of the upstream gene and the start codon of the downstream gene. Using the most accurate gene annotation set, we found a median intergenic distance of 380 bp between upstream genes and the first downstream gene, and a significantly smaller (Wilcoxon rank-sum test: $P = 0.006$) median distance of 250 bp between downstream genes (Fig. 4B). The median distance between monocistronic SL-receiving genes and their downstream neighbors was 2667 bp and the median distance between non-*trans*-spliced genes was 2623 bp (Fig. 4B). The distributions of intergenic distances among these two gene classes were indistinguishable from each other ($P = 0.586$). Across all 20 data sets, the medians ranged from 103 to 749 bp for downstream genes, from 142 to 1028 bp for upstream genes, from 1297 to 3211 bp for monocistronic genes and from 1831 to 3426 bp for non-*trans*-spliced genes (Supplemental Table S2).

An alternative, more precise, metric for identifying genes within operons is the intercistronic distance, defined as the distance between the most abundant polyadenylation [poly(A)] site for the upstream gene to the downstream gene *trans*-splice site (Allen et al. 2011). However, predicting poly(A) sites from our RNA-seq data was limited since the library preparation was not specifically designed to preserve full 3′ end information. Only 1,488,592 reads across all three libraries (0.4% of total reads) contained a poly(A) tail of at least 4 nt coupled with one of 12 polyadenylation signals (Hajarnavis et al. 2004) 16–24 bp upstream

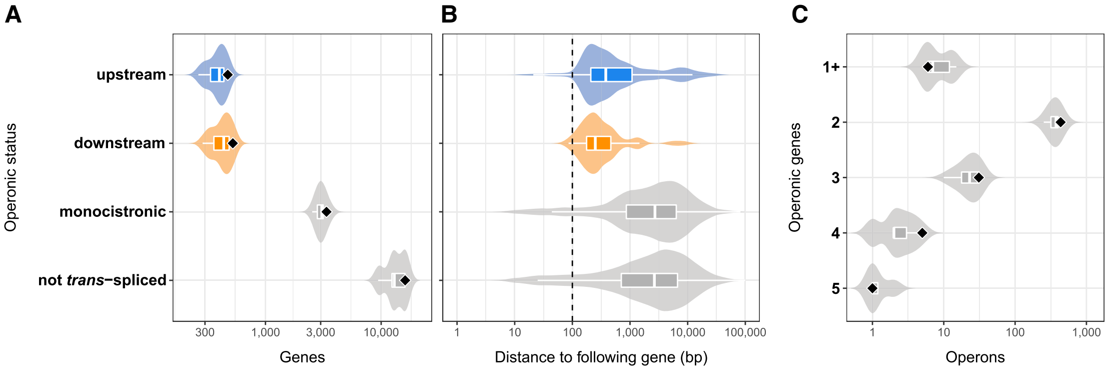

FIGURE 4. Genome-wide identification of *T. spiralis* operons using spliced leader usage patterns. (*A*) Numbers of expressed genes inferred to be upstream or downstream operonic, monocistronic (*trans*-spliced, but not operonic) or not *trans*-spliced. Plots show distributions (density and boxplot) across all 20 data sets; black diamonds indicate values for the most accurate data set (BRAKER + TRINITY, classified with 8 bp minimum match). (*B*) Distributions of physical intercistronic or intergenic distances for the most accurate data set (upstream—distance between first and second gene in operon; downstream—distance between downstream genes in same operon; monocistronic—distance of *trans*-spliced gene to following gene; not *trans*-spliced—distance of non-*trans*-spliced gene to following gene). (*C*) Numbers of inferred operons of particular sizes (1+ operons comprise a single SL2-type spliced leader *trans*-spliced gene without an upstream gene on the same genomic scaffold). Distributions are shown across all 20 data sets; black diamonds indicate values for the most accurate data set.

of the poly(A) tail (Supplemental Table S4). These reads defined 23,392 candidate poly(A) sites across the genome, of which only 8–33 were situated in intercistronic regions (depending on which of the 20 operon prediction sets was used) and at most 200 bp downstream from the 3′ end of the upstream operonic gene. Across all operon annotation sets, the median distances of these few poly(A) sites from the stop codon of their genes ranged between 45 and 146 bp (Supplemental Table S4). Similar results were obtained with a different method (APATrap) (Ye et al. 2018) based on read coverage drop-off at gene ends, which yielded up to 8,441 poly(A) sites, of which a maximum of 124 were situated in intercistronic regions with median distances to the upstream 3′ gene ends of 43–138 bp (Supplemental Table S4). These results indicate that our genome-wide estimate of intercistronic distances, based on the distance between the stop codon of the upstream gene and the start codon of the downstream gene, may be inflated by between 43 and 146 bp, and could thus potentially be reduced to between 60 and 603 bp, assuming this distribution of poly(A) sites. This is consistent with *C. elegans*, where most intercistronic distances are 50–200 bp, and ∼80% of intercistronic regions are less than or equal to 500 bp (Allen et al. 2011). A similar picture emerged when we manually annotated poly(A) sites for our 45 control operons (Supplemental Table S3). We were able to determine the position of the poly(A) sites for 16 intercistronic regions. These gave us a range of intercistronic distances between 74 and 213 bp (median: 124 bp), which are in close agreement with that determined for *C. elegans* operons (median: 129 bp) (Allen et al. 2011).

Previous studies in *C. elegans* have identified the presence of conserved elements, termed Ur motifs, that are implicated in SL2 *trans*-splicing and are expected to be present about 50 bp upstream of the *trans*-splice site (Lasda et al. 2010). Such elements have also been detected in putative *T. spiralis* operons (Pettitt et al. 2014). We found 407,711 occurrences of the Ur (TAYYTT) motif in the *T. spiralis* genome (Supplemental Table S5). Depending on the operon prediction set used, between 48.4% and 64.5% of downstream operonic genes were at most 100 bp downstream from a Ur motif (median distance: 24–49 bp). Among these candidate motifs, TATTTT was most frequent (66.3%–79.5%), followed by TACTTT (13.0%–22.3%), TATCTT (4.2%–9.9%), and TACCTT (0.6%–2.5%) (Supplemental Table S5). There was no significant difference in Ur motif usage between downstream operonic genes that receive either only *Tsp*-SL2, *Tsp*-SL10, or *Tsp*-SL12 ($\chi^2 = 4.58$, df = 6, $P = 0.59$), contrary to our previous observations that these spliced leaders could be functionally differentiated (Fig. 3D; Supplemental Fig. S2).

Although likely incomplete, this is the first time a genome-wide survey of operons has been possible in a distant relative of *C. elegans*, providing us with the opportunity to investigate operon conservation between the two species. The vast majority of *T. spiralis* operons consisted of two genes (Fig. 4C), and the maximum operon length observed in some data sets was five genes (Supplemental Table S2). We predicted five to 15 operons per data set where a bona fide downstream gene (defined by transcripts receiving predominantly SL2-type spliced leaders) had no upstream gene on the same strand and scaffold due to the fragmented nature of the genome assembly (Fig. 4C; Supplemental Table S2). Synteny between our predicted operons and *C. elegans* operons was poor. Although between 85% and 91% of *T. spiralis* operonic genes (depending on operon annotation set used) were assigned a *C. elegans* homolog, *T. spiralis* operonic genes were syntenic with genes in *C. elegans* operons for only 2%–7% of *T. spiralis* operons (Supplemental Table S6). Further, among these syntenic operon pairs, only between 17% and 44% contained the same total numbers of operonic genes in both species. We found no synteny at all among *T. spiralis* operons containing four or more genes, and only a single instance of synteny among three-gene operons (Supplemental Table S6). Thus, it is clear that *C. elegans* operons typically contain more genes than *T. spiralis* operons and that conservation of operons between these species is relatively poor, consistent with the degree of their evolutionary divergence. Among our 20 sets of operon annotations, the BRAKER + Trinity gene set with 10 bp tail overlap during *Tsp*-SL screening and exon-based gene corrections produced the highest levels of synteny (Supplemental Table S6), but even this set supported only 31 syntenic operons compared to our benchmark set of 45 manually curated operons that were fully or partially syntenic with *C. elegans* operons.

Although only a small fraction of *T. spiralis* operonic genes are syntenic with genes in *C. elegans* operons, we do see a strong correlation between the two species in terms of the operonic status of homologous genes. Between 44% and 52% of *C. elegans* genes that are homologous to *T. spiralis* operonic genes are themselves also operonic (Supplemental Table S6). If there was no correlation, we would expect that only 15% of *C. elegans* homologs would be operonic (Allen et al. 2011). These data indicate that there are similar constraints operating in the two nematodes that make certain genes more likely to be found in operons.

Functional characterization of operonic genes

Previous studies have shown that *C. elegans* operons preferentially contain genes encoding molecules critical to basic eukaryotic cell biology (Blumenthal and Gleason 2003; Blumenthal 2004), and RNAi studies show that operons are enriched for genes associated with observable loss-of-function phenotypes (Kamath et al. 2003). Operonic genes are also more likely to be expressed in the hermaphrodite germline compared to monocistronic genes (Reinke and Cutter 2009).

To investigate whether similar patterns exist for *T. spiralis* operonic genes, we examined their association with defined biological processes. We identified 1570 *T. spiralis* genes whose functions are predicted to be associated with germline processes, based on sequence homology with genes expressed in the *C. elegans* germline (see Materials and Methods). Of these, 9.2%–15.5% (depending on the operon prediction set used) corresponded to genes predicted to reside in operons (Supplemental Table S7). This proportion is significantly larger than the genomic background rate of 3.7%–7.0% operonic genes (binomial test: $P < 0.001$), showing that the enrichment of germline expressed genes in operons is likely a general feature of nematode operons.

Additionally, between 6758 and 14,444 genes were mapped to 52,055–69,177 UNIPROT proteins, which yielded 2690–3210 unique GeneOntology annotations (Supplemental Table S8). Across all data sets, 115 unique terms in the biological process ontology were significantly overrepresented among operonic genes. These were simplified to 37 representative terms following clustering by semantic similarity. All these terms represented essential cellular processes, for example RNA modification and metabolism, cellular component organization/localization, protein modification, and cellular respiration (Supplemental Table S8). This is consistent with a view that operonic genes are predominantly involved in fundamental cellular metabolism and regulation of gene expression.

DISCUSSION

We have discovered a remarkable conservation of a secondary structure element shared between the *T. spiralis* spliced leader RNAs, *Tsp*-SL2, SL10, and SL12 and the much better characterized *C. elegans* SL2 RNAs. Genome-wide analysis of *T. spiralis* spliced leader *trans*-splicing shows that presence of this relatively short motif correlates tightly with spliced leader RNA specificity for downstream operonic genes. We have thus uncovered strong evidence for functional conservation between the SL2-type RNAs in these two nematodes. The similar intercistronic distances between *T. spiralis* and *C. elegans* operonic genes provides further evidence for mechanistic conservation, since it indicates that the distance between poly(A) sites and *trans*-splice sites in *T. spiralis* operons is constrained by the interaction between the SL RNP (formed by the SL2-type RNA and associated proteins) and the polyadenylation machinery as it is in *C. elegans* (Evans et al. 2001).

However we observe differences in SL1 versus SL2 *trans*-splicing in the two nematodes. In *C. elegans*, SL2 spliced leaders are almost never recruited to pre-mRNAs derived from monocistronic genes or the first genes in operons (Tourasse et al. 2017), whereas in *T. spiralis* ∼10% of such pre-mRNAs are *trans*-spliced to SL2-type spliced leaders. This is likely explained by *C. elegans* SL1 RNA being more abundant than SL2 RNAs (supported by the fact that in mutants which lack SL1 RNA, SL2s are *trans*-spliced to outrons of pre-mRNAs [Ferguson et al. 1996]), whereas in *T. spiralis* the expression levels of the different SL RNAs are likely in the same range.

We also see differences between the three *T. spiralis* SL2 RNAs: *Tsp*-SL2 appears to be more effective at *trans*-splicing to outrons than *Tsp*-SL10 and *Tsp*-SL12, for instance. A simple explanation would be that it is the more highly expressed of the SL2-type SL RNAs. However, expression levels cannot explain all the differences between spliced leaders, since we would then expect *Tsp*-SL2 to also be the dominant spliced leader *trans*-spliced to downstream operonic pre-mRNAs, whereas we find that *Tsp*-SL12 is the most frequently selected. This suggests that there are subtle functional differences between the three SL2-type SL RNAs for which we cannot currently account.

The same strict sequence conservation of stem–loop 3 was found in other Clade I nematode SL RNAs, which indicates that SL2-type *trans*-splicing is broadly distributed across this taxon. Revisiting our own work on nematode spliced leader *trans*-splicing and operons reinforces this conclusion. All three *T. muris* SL2-type spliced leaders are *trans*-spliced to mRNAs expressed by putative downstream operonic genes in this nematode (Pettitt et al. 2014). More significantly, heterologous expression of one of the *P. punctatus* putative SL2-type spliced leaders, *Ppu*-SL2, in *C. elegans* resulted in the *trans*-splicing onto mRNAs exclusively derived from downstream operonic genes (Harrison et al. 2010). Finally, a *T. spiralis* intercistronic region, which includes a credible Ur element, serves as an exclusive SL2 RNA substrate in *C. elegans* (Pettitt et al. 2014).

The strict sequence conservation of the stem–loop 3 motif across multiple nematode taxa raises the possibility that SL2-type *trans*-splicing arose prior to, or during early nematode evolution. However, we cannot exclude the possibility that the sequence conservation we observe is due to convergent evolution, with SL2-type *trans*-splicing having independently arisen in the lineages leading to Clade V and Clade I nematodes. We think that this second scenario is less likely, as the convergent evolution of the stem–loop 3 motif would need to be accompanied by parallel coevolution of the interaction with the polyadenylation machinery (Evans and Blumenthal 2000; Evans et al. 2001; Lasda et al. 2010). Expanding our analysis into the other nematode clades could address this uncertainty. We have shown that two Clade IV nematode species have putative SL2-type spliced leaders (Supplemental Fig. S1). However, we failed to detect any credible candidates from the Spirurina clade and we have no molecular information about spliced leader *trans*-splicing in the least well-characterized nematode clade, the Enoplia (Clade II). A key future investigation will be to determine whether there is a single phylogenetic origin for SL2-type polycistronic processing, and whether it predates the foundation of the Nematoda.

SL2 *trans*-splicing is a diagnostic feature of operons in *C. elegans* and other members of the Rhabditina. Our discovery of SL2-type *trans*-splicing in *T. spiralis* allowed us to reliably define a minimal set of operons in this nematode for the first time, facilitating the comparison of the operon repertoires of two distantly related nematodes. Previous interspecies investigations into operonic gene function, patterns of operon synteny, and operon reorganization have focused on nematodes that share many derived molecular and cellular features (Guiliano and Blaxter 2006; Ghedin et al. 2007; Qian and Zhang 2008; Cutter and Agrawal 2010; Sinha et al. 2014). Comparison between *C. elegans* and *T. spiralis* allows a deeper view of the processes that have shaped nematode operons.

Although the proportion of *T. spiralis* genes we identify as operonic (~4%) is lower than in *C. elegans* (15%), the genes that are found in operons show the same general pattern, encoding basic cellular functions, and showing enrichment for germline expression. This supports the hypothesis that operons facilitate recovery from developmental arrest, a life cycle strategy that is common throughout the Nematoda (Zaslaver et al. 2011). Similarly, the size distribution of operons in *T. spiralis*, in terms of gene number per operon, resembles those for *C. elegans* and *C. briggsae* (Uyar et al. 2012). Taken together, the data suggest that the dynamics of operon formation and maintenance are broadly similar across the Nematoda. Consistent with comparisons across shorter evolutionary distances (Ghedin et al. 2007; Sinha et al. 2014), we don't find strong evidence for extensive syntenic conservation of operons when comparing *T. spiralis* to *C. elegans*. Our data indicate that extensive reorganization of operons have occured since the separation of the two lineages during the radiation of the Nematoda.

Finally, the discovery that SL2-type *trans*-splicing is a broadly distributed nematode trait should lead to dramatic improvements to nematode genome annotations. Conservation of the stem–loop 3 motif can be used to identify SL2-type spliced leaders. Provided that the spliced leader sequence can be diagnostically associated with this motif (which is the case for all instances that we have analyzed, but a priori need not be), then identification of transcripts *trans*-spliced to these spliced leaders would provide a definitive means to identify operons and operonic genes, an approach that has previously been confined to *C. elegans* and its close relatives, but which should see broad application in multiple nematode genomes, offering an avenue for greatly improving their genome annotations.

MATERIALS AND METHODS

Genome-wide identification of spliced-leader *trans*-splicing events

RNA-seq library preparation and sequencing

Three independent sets of five outbred ICR/CD1 female mice (10–12 wk old) were infected with 400–500 *Trichinella spiralis* L1 muscle larvae by oral gavage. Approximately 100,000 muscle larvae were recovered four months post infection using the pepsin/HCL digestion method described by Blair (1983). The larvae were pooled for RNA extraction with the PureLink RNA mini kit (Ambion by Life Technologies) following manufacturer's instructions and including an on-column DNA removal step with PureLink DNase. Three unstranded Illumina TruSeq mRNA V2 libraries were prepared and sequenced on a single lane of an Illumina HiSeq 2000/2500 instrument in 101 bp paired-end mode at The Genome Analysis Centre (TGAC) in Norwich, UK (now The Earlham Institute).

Data quality was assessed using FASTQC 0.11.3 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), followed by trimming of Illumina adapter sequences and bases with phred quality score below 20 using TRIM_GALORE 0.4.0 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Read pairs containing at least one read shorter than 30 bp after trimming were discarded.

Identification of reads with Tsp-SL sequences

Reads with evidence of spliced leader *trans*-splicing were identified using a strategy modified from published pipelines designed for *C. elegans* (Tourasse et al. 2017; Yague-Sanz and Hermand 2018). Read pairs were aligned to the *Trichinella spiralis* 3.7.1 reference genome assembly (NCBI accession GCA_000181795.2; BioProject PRJNA12603) using HISAT 2.1.0 (Kim et al. 2015) with enforced end-to-end alignment of reads and concordant alignment of read pairs. Reads containing a spliced leader are expected not to align end-to-end due to the spliced-leader overhang, whereas their mate read is expected to align end-to-end (Yague-Sanz and Hermand 2018). Since the RNA-seq libraries were unstranded, spliced leader tags can occur on R1 or R2 reads; for ease of downstream processing, we pseudo-stranded identified candidate read pairs such that all unaligned reads were designated as R1 and all aligned mates as R2, using SAMTools 1.6 (Li et al. 2009). These candidate read pairs were then screened for 15 known *Tsp*-SL sequences (Pettitt et al. 2008) at their 5′ ends using Cutadapt 1.15 (Martin 2011) with a minimum perfect match of 10 bp in order to ensure unambiguous *Tsp*-SL assignment, and with a minimum read length of 20 bp after trimming all matching *Tsp*-SL bases. An alternative screening was undertaken with a minimum perfect match of 8 bp; while this recovered a larger number of *Tsp*-SL reads, *Tsp*-SL13, *Tsp*-SL14, and *Tsp*-SL15 cannot be distinguished reliably (Pettitt et al. 2008).

De novo genome reannotation

The genome was reannotated de novo using the concordant read-pair alignments generated from the three RNA-seq libraries. Transcript sequences were assembled in genome-guided Trinity 2.5.0 (Grabherr et al. 2011), ORFs were extracted using Transdecoder 5.3.0 (Haas et al. 2013), and translated protein sequences were clustered at 100% similarity using CD-HIT 4.7 (Fu et al. 2012). This nonredundant set of proteins was then used alongside the RNA-seq alignments to generate AUGUSTUS-based gene predictions using Braker 2.1.0 (Hoff et al. 2015) after soft-masking repetitive sequences with RepeatMasker 4.0.1 (Chen 2004). For comparison, two further sets of annotations were generated from the RNA-seq alignments only, using Braker or StringTie 1.3.4d (Pertea et al. 2015). A final set was generated by merging these three sets to a nonredundant set of loci using GffCompare 0.10.6 (https://ccb.jhu.edu/software/stringtie/gffcompare.shtml). Together with the published reference annotations, we thus worked with five annotation sets in total. We examined annotation quality with BUSCO 3.0.2 (Simão et al. 2015) using the nematode ortholog set (3131 orthologs).

Tsp-SL read quantification

The screened and trimmed *Tsp*-SL read pairs were aligned back to the genome using HISAT2, enforcing end-to-end alignment. Aligned R1 reads were extracted using SAMTools 1.6 (Li et al. 2009) and quantified against the reference gene annotations as well as our four de novo gene annotation sets using featureCounts 1.6.1 (Liao et al. 2014). Read counts were normalized by library size and expressed as counts-per-million (CPM) (Allen et al. 2011). Since *Tsp*-SL reads are expected to align to the first exon of the trans-spliced gene, incorrect gene annotations can be identified via internal exons that receive *Tsp*-SL reads. To explore the extent of this issue, we generated a nonredundant set of exon annotations across isoforms using BEDTools Merge 2.26 (Quinlan and Hall 2010) and requantified the *Tsp*-SL reads at the exon level. These exon counts were then processed in R 3.5.1 (R Core team 2013) to split gene annotations at internal exons that received a distinct peak of at least four counts (and at least one count per library) compared to neighboring exons. We assumed that each *Tsp*-SL-receiving exon demarcates the beginning of a new gene and corrected the gene annotations accordingly.

In silico prediction of *T. spiralis* spliced leader RNA secondary structure and identification of the conserved stem–loop 3 motif in other nematode SL RNAs

We predicted the secondary structure of SL RNA sequences in *T. spiralis* (Pettitt et al. 2008) and previously identified SL RNAs from the other Clade I nematodes *Trichuris muris* and *Prionchulus punctatus* (Harrison et al. 2010; Pettitt et al. 2014). On the basis of the conserved stem–loop 3 and associated Sm binding site, we used Infernal 1.1.3 (Nawrocki and Eddy 2013) to perform covariance model-based identification of SL2-type RNA from the genomes of the Clade I nematodes, *Romanomermis culicivorax* (GCA_001039655.1), *Trichinella pseudospiralis* (GCA_001447575.1), and *Soboliphyme baturini* (GCA_900618415.1); and the Clade IV nematodes, *Bursaphelenchus xylophilus* (GCA_000231135.1) and *Panagrellus redivivus* (GCA_000341325.1). We used the software Mfold 3.6 (Zuker 2003) (with the default folding conditions and the constraint that putative Sm-binding sites were required to be single stranded) to predict the secondary structures of newly identified SL RNAs.

Identification of functional specialization among *Tsp*-SL types

To investigate functional specializations consistent with the secondary structure predictions, we used multivariate methods to cluster *Tsp*-SL types by similarity of gene sets targeted by each type, using the adegenet 2.1.1 (Jombart 2008) and DeSeq2 1.22.1 (Love et al. 2014) R packages. The gene counts obtained for each *Tsp*-SL type among our three libraries (see above) were transformed into a single presence/absence matrix, where a gene was present in a sample (*Tsp*-SL type) if it received at least one read. Jaccard distances among samples were estimated from this matrix, projected onto two-dimensional space using metric multidimensional scaling (MDS) and hierarchically clustered using Ward's criterion. Discrimination between evident groups of samples was then explored using linear discriminant analysis of PCA (DAPC) (Jombart 2008). The best number of principal components to retain was identified using the cross-validation function within DAPC. The same analyses were also carried out on variance-stabilized normalized read counts using PCA instead of MDS (Love et al. 2014).

The presence of any distinct functional classes of spliced leader RNAs would suggest the possibility that different classes could have distinct roles in resolving monocistronic/upstream operonic versus downstream operonic mRNAs, that is, the SL1- and SL2-type found in *C. elegans* (Blumenthal 2012). To investigate this, we examined whether genes within a known set of 45 operons conserved between *C. elegans* and *T. spiralis* (Pettitt et al. 2014) differ systematically in the types of *Tsp*-SL they receive. Protein sequences of operonic genes in *C. elegans* were extracted from the WS247 genome release and orthologs were identified in the WS247 *T. spiralis* proteome using BLASTP 2.2.7 (Camacho et al. 2009) with an E-value cutoff of 1×10^{-5}. The best match (if any) for each gene was then checked against corresponding matches to other constituents in the *C. elegans* operon. Operon conservation was established if the *T. spiralis* orthologs were on the same scaffold, the same strand, within 5 kbp of each other, and in the same orientation as the *C. elegans* operon genes. If the BLAST search for two different *C. elegans* operon constituents returned a single *T. spiralis* gene, but the homology matches were to different parts of the gene and in the correct orientation, this was also annotated as a conserved operon. In this case, this was likely an operon that had been incorrectly annotated as a single gene (Pettitt et al. 2014), as we have previously observed. Of the 45 operons, three were previously identified (Pettitt et al. 2014). For each conserved operonic *T. spiralis* gene, we annotated the *Tsp*-SL type(s) that the gene receives using IGV 2.6.3 (Robinson et al. 2011). We also manually annotated the location of experimentally confirmed or predicted polyadenylation signals, allowing us to define the intercistronic regions within each operon.

Genome-wide prediction of operons

Identification of operonic genes and intercistronic distances

We predicted operons genome-wide on the basis of the alignment locations of SL2-type *Tsp*-SL reads. Normalized read counts for each *Tsp*-SL were summarized across the three replicate libraries using the geometric mean while allowing for zero counts in one library. These mean counts were then added among all SL1-type and SL2-type *Tsp*-SLs, and an SL2:SL1-type ratio was computed for each gene. Uninterrupted runs of at least one predominantly SL2-type receiving gene (SL2:SL1 > 2) along each scaffold and strand of the reference genome were designated as "downstream operonic" genes (Blumenthal et al. 2002). The immediately adjacent gene upstream of each tract of downstream genes was designated "upstream operonic," irrespective of whether it was *trans*-spliced (SL2:SL1 ≤ 2) or not. All other spliced leader *trans*-spliced genes were designated as "monocistronic."

An important consistency check for the predicted operons is that intercistronic distances (defined by the distance between the polyadenylation site of the upstream gene and the *trans*-splicing acceptor site of the downstream gene) should be considerably reduced compared to distances between nonoperonic genes. In *C. elegans*, intercistronic distances have a median value of 129 bp (Allen et al. 2011), indicating tight organization of operonic genes. Since poly(A) sites are not annotated, we calculated intergenic distances in our five *T. spiralis* genome annotations using the boundaries of the gene annotations (defined by start/stop codons). To explore the extent to which these intergenic distances may be inflated compared to intercistronic distances, we attempted to predict poly(A) sites from the RNA-seq data. Since the libraries were generated with conventional protocols that are unlikely to preserve full 3′ information compared to specialized 3′-targeted library protocols (e.g., Welch et al. 2015; Routh 2019), we applied two different methods. First, we used the coverage-based method of reconstructing 3′-UTRs and poly(A) sites implemented in APAtrap (Ye et al. 2018) in "short 3′-UTR" mode, requiring read coverage of 10 and distance between poly(A) sites of 100 bp. Second, we screened the RNA-seq reads for poly(A) tails using Cutadapt 1.15 (Martin 2011), trimming a tail of at least 4 A nucleotides and allowing an error rate of 0.167 for longer tails. All identified reads of at least 30 bp length after poly(A) trimming were filtered further by the presence of the quintessential polyadenylation signal AAUAAA (or 12 alternative signals known in *C. elegans*) (Hajarnavis et al. 2004) starting 16–24 bp upstream of the poly(A) tail. The filtered reads were then aligned against the *T. spiralis* genome using BLASTN 2.6.0 (Camacho et al. 2009) and the genomic positions of the 3′ ends of the reads were extracted and processed in R. We defined a poly(A) site as a collection of at least three reads whose 3′ alignment positions fall within a 20 bp window, and consecutive poly(A) sites are at least 100 bp apart. The position of each predicted poly(A) site was taken to be the median position among all reads within that site. The distances of the predicted poly(A) sites (from both methods) to the 3′ end of the nearest operonic gene were then obtained with BEDTools Closest (Quinlan and Hall 2010). We summarized those distances for the poly(A) site that resided in the intercistronic space up to 200 bp downstream from the 3′ end of the upstream gene.

Another consistency check is that genes undergoing SL2-type *trans*-splicing should have the U-rich (Ur) element motif present about 50 bp upstream of the *trans*-splice site (Lasda et al. 2010). We scanned the *T. spiralis* genome for the Ur motif TAYYTT using EMBOSS Dreg 6.5.7.0 (Rice et al. 2000) and identified those motifs that lie up to 100 bp upstream of a downstream operonic gene using BEDTools Closest. We also examined whether particular Ur motif types (e.g., TATTTT or TACCTT) are associated with particular SL2-types at downstream operonic genes.

Operon synteny with C. elegans

We examined the degree of synteny between all predicted *T. spiralis* operons and *C. elegans* operons, going beyond our original set of 45 benchmark operons. The exonic nucleotide sequence of each operonic gene was extracted using GffRead 0.9.9 (Trapnell et al. 2010) and homologous *C. elegans* proteins (PRJNA13758. WBPS14) were retrieved using BLASTX 2.6.0 (Camacho et al. 2009). The single hit with the lowest E-value was retained and the *C. elegans* gene identifiers were extracted. These identifiers were then searched against operon definitions contained in the *C. elegans* genome annotations (PRJNA13758.WBPS14). Since we observed that these genome annotations erroneously assign 135 genes to more than one operon, we manually assigned the correct operon to these genes using the WORMBASE JBROWSE genome browser. For each *T. spiralis* operon we examined whether the homologous *C. elegans* genes were also operonic, in the same gene order and residing in a *C. elegans* operon with the same numbers of genes as the *T. spiralis* operon.

Functional enrichment of operonic genes

We carried out basic functional annotation of operonic genes to examine whether these genes may be enriched for particular biological functions. In *C. elegans*, it has been shown that 38% of genes expressed in the germline are operonic, whereas only 15% of all genes in the genome are operonic (Reinke and Cutter 2009). Accordingly, we first tested the hypothesis whether *T. spiralis* genes involved in germline processes are more frequently located in operons than expected from the genomic background rate of operonic organization. *C. elegans* germline gene names were extracted from Reinke et al. (2004), translated to *C. elegans* gene IDs using WormBase ParaSite BioMart (WBPS12; WS267) and corresponding protein sequences were extracted from the WormBase *C. elegans* reference proteome. *T. spiralis* homologs for these proteins were obtained using TBLASTN 2.9.0 against the *T. spiralis* genome with an e-value cutoff of 1×10^{-5}, and tested for overlap with our predicted *T. spiralis* operons using BEDTools Intersect 2.26 (Quinlan and Hall 2010). The proportion of orthologs overlapping operons was tested for deviation from the expected overall proportion of operonic genes using binomial and χ^2 tests carried out in R.

Additionally, we carried out GeneOntology annotation and enrichment tests for operonic genes. Transcript nucleotide sequences based on each of the five sets of annotations were extracted from the reference genome using GffRead 0.9.9 (Trapnell et al. 2010). The sequences were queried against the UniProt swissprot and Nematoda protein databases using BLASTX 2.6.0 (Camacho et al. 2009) and retaining up to 10 hits with an E-value cutoff of 1×10^{-3}. Gene Ontology terms associated with matching UniProt accessions were retrieved from the Gene Ontology Annotation (GOA) database (Huntley et al. 2014). Enrichment tests were carried out with the R package GOfuncR 1.2 (https://www.bioconductor.org/packages/release/bioc/html/GOfuncR.html), comparing the biological process ontology annotations of all operonic genes against the background annotations of the whole genome using the hypergeometric test and retaining significantly enriched annotations at an FDR-corrected *P*-value threshold of $q \le 0.05$. Significantly enriched annotations were pooled across all data sets and semantically clustered at a similarity threshold of 0.4 (SimRel measure) using REVIGO (Supek et al. 2011).

DATA DEPOSITION

RNA-seq reads for the three *T. spiralis* libraries are available from the NCBI SRA database, accession numbers SRR8327925–SRR8327927 (bioproject PRJNA510020). All scripts used for these analyses are available at https://github.com/glewgun/Wenzel/ or upon request.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work was supported by the Biotechnology and Biological Sciences Research Council (BB/J007137/1 to J.P., B.M., and B. C., and BB/T002859/1 to B.M. and J.P.) and an EASTBIO Biotechnology and Biological Sciences Research Council PhD Studentship (BB/M010996/1 to C.J.). The authors would like to acknowledge the support of the Maxwell computer cluster funded by the University of Aberdeen.

Author contributions: B.C. provided the biological material; J. P., B.C., and B.M. acquired funding for the project; B.M., J.P., and B.C. conceived the research and managed and coordinated the research activity; M.W. and C.J. designed and implemented the computational analysis; M.W., B.M., J.P., and B.C. analyzed the data; M.W. and J.P. prepared the figures and tables. J.P., M.W., B.C., and B.M. wrote the manuscript.

Received May 18, 2020; accepted August 26, 2020.

REFERENCES

Allen MA, Hillier LW, Waterston RH, Blumenthal T. 2011. A global analysis of *C. elegans trans*-splicing. *Genome Res* **21:** 255–264. doi:10.1101/gr.113811.110

Blair LS. 1983. Laboratory techniques. In *Trichinella and trichinosis* (ed. Campbell WC), pp. 563–570. Springer US, Boston, MA.

Blumenthal T. 2004. Operons in eukaryotes. *Brief Funct Genomic Proteomic* **3:** 199–211. doi:10.1093/bfgp/3.3.199

Blumenthal T. 2012. *Trans*-splicing and operons in *C. elegans*. *WormBook* 1–11. doi:10.1895/wormbook.1.5.2

Blumenthal T, Gleason KS. 2003. *Caenorhabditis elegans* operons: form and function. *Nat Rev Genet* **4:** 112–120. doi:10.1038/nrg995

Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M, et al. 2002. A global analysis of *Caenorhabditis elegans* operons. *Nature* **417:** 851–854. doi:10.1038/nature00831

Blumenthal T, Davis P, Garrido-Lecca A. 2015. Operon and non-operon gene clusters in the *C. elegans* genome. *WormBook* 1–20. doi:10.1895/wormbook.1.175.1

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10:** 421. doi:10.1186/1471-2105-10-421

Chen N. 2004. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **5:** 4–10. doi:10.1002/0471250953.bi0410s05

Cutter AD, Agrawal AF. 2010. The evolutionary dynamics of operon distributions in eukaryote genomes. *Genetics* **185:** 685–693. doi:10.1534/genetics.110.115766

Danks GB, Raasholm M, Campsteijn C, Long AM, Manak JR, Lenhard B, Thompson EM. 2015. *Trans*-splicing and operons in metazoans: translational control in maternally regulated development and recovery from growth arrest. *Mol Biol Evol* **32:** 585–599. doi:10.1093/molbev/msu336

Douris V, Telford MJ, Averof M. 2010. Evidence for multiple independent origins of *trans*-splicing in Metazoa. *Mol Biol Evol* **27:** 684–693. doi:10.1093/molbev/msp286

Evans D, Blumenthal T. 2000. *trans*-splicing of polycistronic *Caenorhabditis elegans* pre-mRNAs: analysis of the SL2 RNA. *Mol Cell Biol* **20:** 6659–6667. doi:10.1128/MCB.20.18.6659-6667.2000

Evans D, Zorio D, MacMorris M, Winter CE, Lea K, Blumenthal T. 1997. Operons and SL2 *trans*-splicing exist in nematodes outside the genus *Caenorhabditis*. *Proc Natl Acad Sci* **94:** 9751–9756. doi:10.1073/pnas.94.18.9751

Evans D, Perez I, MacMorris M, Leake D, Wilusz CJ, Blumenthal T. 2001. A complex containing CstF-64 and the SL2 snRNP connects mRNA 3′ end formation and *trans*-splicing in *C. elegans* operons. *Genes Dev* **15:** 2562–2571. doi:10.1101/gad.920501

Ferguson KC, Heid PJ, Rothman JH. 1996. The SL1 *trans*-spliced leader RNA performs an essential embryonic function in *Caenorhabditis elegans* that can also be supplied by SL2 RNA. *Genes Dev* **10:** 1543–1556. doi:10.1101/gad.10.12.1543

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28:** 3150–3152. doi:10.1093/bioinformatics/bts565

Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, Allen JE, Delcher AL, Guiliano DB, Miranda-Saavedra D, et al. 2007. Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317:** 1756–1760. doi:10.1126/science.1145406

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29:** 644. doi:10.1038/nbt.1883

Guiliano DB, Blaxter ML. 2006. Operon conservation and the evolution of *trans*-splicing in the phylum nematoda. *PLoS Genet* **2:** e198. doi:10.1371/journal.pgen.0020198

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8:** 1494. doi:10.1038/nprot.2013.084

Haenni S, Sharpe HE, Gravato Nobre M, Zechner K, Browne C, Hodgkin J, Furger A. 2009. Regulation of transcription termination in the nematode *Caenorhabditis elegans*. *Nucleic Acids Res* **37:** 6723–6736. doi:10.1093/nar/gkp744

Hajarnavis A, Korf I, Durbin R. 2004. A probabilistic model of 3′ end formation in *Caenorhabditis elegans*. *Nucleic Acids Res* **32:** 3392–3399. doi:10.1093/nar/gkh656

Harrison N, Kalbfleisch A, Connolly B, Pettitt J, Müller B. 2010. SL2-like spliced leader RNAs in the basal nematode *Prionchulus punctatus*: new insight into the evolution of nematode SL2 RNAs. *RNA* **16:** 1500–1507. doi:10.1261/rna.2155010

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2015. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32:** 767–769. doi:10.1093/bioinformatics/btv661

Huang XY, Hirsh D. 1989. A second *trans*-spliced RNA leader sequence in the nematode *Caenorhabditis elegans*. *Proc Natl Acad Sci* **86:** 8640–8644. doi:10.1073/pnas.86.22.8640

Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'donovan C. 2014. The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Res* **43:** D1057–D1063. doi:10.1093/nar/gku1113

Johnston C. 2018. *Genome-wide analysis of spliced leader trans-splicing in the nematode Trichinella spiralis*. PhD thesis, University of Aberdeen.

Jombart T. 2008. *adegenet*: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24:** 1403–1405. doi:10.1093/bioinformatics/btn129

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, et al. 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421:** 231–237. doi:10.1038/nature01278

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12:** 357. doi:10.1038/nmeth.3317

Krause M, Hirsh D. 1987. A *trans*-spliced leader sequence on actin mRNA in *C. elegans*. *Cell* **49:** 753–761. doi:10.1016/0092-8674(87)90613-1

Lasda EL, Blumenthal T. 2011. *Trans*-splicing. *Wiley Interdiscip Rev RNA* **2:** 417–434. doi:10.1002/wrna.71

Lasda EL, Allen MA, Blumenthal T. 2010. Polycistronic pre-mRNA processing in vitro: snRNP and pre-mRNA role reversal in *trans*-splicing. *Genes Dev* **24:** 1645–1658. doi:10.1101/gad.1940010

Lee K-Z, Sommer RJ. 2003. Operon structure and *trans*-splicing in the nematode *Pristionchus pacificus*. *Mol Biol Evol* **20:** 2097–2103. doi:10.1093/molbev/msg225

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930. doi:10.1093/bioinformatics/btt656

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550. doi:10.1186/s13059-014-0550-8

MacMorris M, Kumar M, Lasda E, Larsen A, Kraemer B, Blumenthal T. 2007. A novel family of *C. elegans* snRNPs contains proteins associated with *trans*-splicing. *RNA* **13:** 511–520. doi:10.1261/rna.426707

Maroney PA, Denker JA, Darzynkiewicz E, Laneve R, Nilsen TW. 1995. Most mRNAs in the nematode *Ascaris lumbricoides* are *trans*-spliced: a role for spliced leader addition in translational efficiency. *RNA* **1:** 714–723.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal* **17:** 10. doi:10.14806/ej.17.1.200

Mitreva M, Jasmer DP, Zarlenga DS, Wang Z, Abubucker S, Martin J, Taylor CM, Yin Y, Fulton L, Minx P, et al. 2011. The draft genome of the parasitic nematode *Trichinella spiralis*. *Nat Genet* **43:** 228–235. doi:10.1038/ng.769

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29:** 2933–2935. doi:10.1093/bioinformatics/btt509

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33:** 290. doi:10.1038/nbt.3122

Pettitt J, Müller B, Stansfield I, Connolly B. 2008. Spliced leader *trans*-splicing in the nematode *Trichinella spiralis* uses highly polymorphic, noncanonical spliced leaders. *RNA* **14:** 760–770. doi:10.1261/rna.948008

Pettitt J, Philippe L, Sarkar D, Johnston C, Gothe HJ, Massie D, Connolly B, Müller B. 2014. Operons are a conserved feature of nematode genomes. *Genetics* **197:** 1201–1211. doi:10.1534/genetics.114.162875

Qian W, Zhang J. 2008. Evolutionary dynamics of nematode operons: easy come, slow go. *Genome Res* **18:** 412–421. doi:10.1101/gr.7112608

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

R Core team. 2013. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Reinke V, Cutter AD. 2009. Germline expression influences operon organization in the *Caenorhabditis elegans* genome. *Genetics* **181:** 1219–1228. doi:10.1534/genetics.108.099283

Reinke V, San Gil I, Ward S, Kazmer K. 2004. Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. *Development* **131:** 311–323. doi:10.1242/dev.00914

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16:** 276–277. doi:10.1016/S0168-9525(00)02024-2

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26. doi:10.1038/nbt.1754

Routh A. 2019. DPAC: a tool for differential poly(A)-cluster usage from poly(A)-targeted RNAseq data. *G3* **9:** 1825–1830.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31:** 3210–3212. doi:10.1093/bioinformatics/btv351

Sinha A, Langnick C, Sommer RJ, Dieterich C. 2014. Genome-wide analysis of *trans*-splicing in the nematode *Pristionchus pacificus* unravels conserved gene functions for germline and dauer development in divergent operons. *RNA* **20:** 1386–1397. doi:10.1261/rna.041954.113

Spieth J, Brooke G, Kuersten S, Lea K, Blumenthal T. 1993. Operons in *C. elegans*: polycistronic mRNA precursors are processed by *trans*-splicing of SL2 to downstream coding regions. *Cell* **73:** 521–532. doi:10.1016/0092-8674(93)90139-H

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6:** e21800. doi:10.1371/journal.pone.0021800

Tourasse NJ, Millet JRM, Dupuy D. 2017. Quantitative RNA-seq meta-analysis of alternative exon usage in *C. elegans*. *Genome Res* **27:** 2120–2128. doi:10.1101/gr.224626.117

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511. doi:10.1038/nbt.1621

Uyar B, Chu JSC, Vergara IA, Chua SY, Jones MR, Wong T, Baillie DL, Chen N. 2012. RNA-seq analysis of the *C. briggsae* transcriptome. *Genome Res* **22:** 1567–1580. doi:10.1101/gr.134601.111

Welch JD, Slevin MK, Tatomer DC, Duronio RJ, Prins JF, Marzluff WF. 2015. EnD-Seq and AppEnD: sequencing 3′ ends to identify nontemplated tails and degradation intermediates. *RNA* **21:** 1375–1389. doi:10.1261/rna.048785.114

Yague-Sanz C, Hermand D. 2018. SL-quant: a fast and flexible pipeline to quantify spliced leader *trans*-splicing events from RNA-seq data. *Gigascience* **7:** giy084. doi:10.1093/gigascience/giy084

Yang Y-F, Zhang X, Ma X, Zhao T, Sun Q, Huan Q, Wu S, Du Z, Qian W. 2017. *Trans*-splicing enhances translational efficiency in *C. elegans*. *Genome Res* **27:** 1525–1535. doi:10.1101/gr.202150.115

Ye C, Long Y, Ji G, Li QQ, Wu X. 2018. APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics* **34:** 1841–1849. doi:10.1093/bioinformatics/bty029

Zaslaver A, Baugh LR, Sternberg PW. 2011. Metazoan operons accelerate recovery from growth-arrested states. *Cell* **145:** 981–992. doi:10.1016/j.cell.2011.05.013

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406–3415. doi:10.1093/nar/gkg595