SLIDR and SLOPPR: Flexible identification of spliced leader *trans*-splicing and prediction of eukaryotic operons from RNA-Seq data

Marius A. Wenzel[1*], Berndt Mueller[2] and Jonathan Pettitt[2]

December 18, 2020

[1] School of Biological Sciences, University of Aberdeen, Zoology Building, Tillydrone Ave, Aberdeen AB24 2TZ, UK;

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

*corresponding author

Abstract

Background: Spliced leader (SL) *trans*-splicing replaces the 5' ends of pre-mRNAs with the spliced leader, an exon derived from a specialised non-coding RNA originating from a different genomic location. This process is essential for resolving polycistronic pre-mRNAs produced by eukaryotic operons into monocistronic transcripts. SL *trans*-splicing and operons have independently evolved multiple times throughout Eukarya, but our understanding of these phenomena is limited to only a few well-characterised organisms, most notably *C. elegans* and trypanosomes. The primary barrier to systematic discovery and characterisation of SL *trans*-splicing and operons is the lack of computational tools for exploiting the surge of transcriptomic and genomic resources for a wide range of eukaryotes.

Results: Here we present two novel pipelines that automate the discovery of SLs and the prediction of operons in eukaryotic genomes from RNA-Seq data. SLIDR assembles putative SLs from 5' read tails present after read alignment to a reference genome or transcriptome, which are then verified by interrogation of sequence motifs expected in *bona fide* SL RNA molecules. SLOPPR identifies RNA-Seq reads that contain a given 5' SL sequence, quantifies genome-wide SL *trans*-splicing events and predicts operons via distinct patterns of SL *trans*-splicing events across adjacent genes. We tested both pipelines with organisms known to carry out SL *trans*-splicing and organise their genes into operons, and demonstrate that 1) SLIDR correctly identifies known SLs and often discovers novel SL variants; 2) SLOPPR correctly identifies functionally specialised SLs, correctly predicts known operons and detects plausible novel operons.

Conclusions: SLIDR and SLOPPR are flexible tools that will accelerate research into the evolutionary dynamics of SL *trans*-splicing and operons throughout Eukarya, and improve gene discovery and annotation for a wide-range of eukaryotic genomes. Both pipelines are implemented in Bash and R and are built upon readily available software commonly installed on most bioinformatics servers. Biological insight can be gleaned even from sparse, low-coverage datasets, implying that an untapped wealth of information can be derived from existing RNA-Seq datasets as well as from novel full-isoform sequencing protocols as they become more widely available.

Keywords: spliced-leader *trans*-splicing, eukaryotic operons, polycistronic RNA processing, RNA-Seq, genome annotation, chimeric reads, 5' UTR

Background

Spliced leader (SL) *trans*-splicing is a eukaryotic post-transcriptional RNA modification whereby the 5' end of a pre-mRNA receives a short "leader" exon from a non-coding RNA molecule that originates from elsewhere in the genome [1, 2]. This mechanism was first discovered in trypanosomes [3] and has received much attention as a potential target for diagnosis and control of a range of medically and agriculturally important pathogens [1, 4, 5]. SL *trans*-splicing is broadly distributed among many eukaryotic groups, for example euglenozoans, dinoflagellates, cnidarians, ctenophores, platyhelminthes, tunicates and nematodes, but is absent from vertebrates, insects, plants and fungi [2]. Its phylogenetic distribution and rich molecular diversity suggest that it has evolved independently many times throughout eukaryote evolution [6–9].

One clear biological function of SL *trans*-splicing is the processing of polycistronic pre-mRNAs generated by eukaryotic operons [2]. In contrast to prokaryotes, where such transcripts can be translated immediately as they are transcribed, a key complication for eukaryotic operons is that nuclear polycistronic transcripts must be resolved into independent, 5'-capped monocistronic transcripts for translation in the cytoplasm [10]. The *trans*-splicing machinery coordinates cleavage of polycistronic pre-mRNA and provides the essential cap to initially un-capped pre-mRNAs [11, 12]. This process is best characterised in the nematodes, largely, but not exclusively due to work on *C. elegans*, which possesses two types of SL [13]: SL1, which is added to mRNAs derived from the first gene in operons and monocistronic genes; and SL2, which is added to mRNAs arising from genes downstream in operons and thus specialises in resolving polycistronic pre-mRNAs [11–13].

The same SL2-type specialisation of some SLs for resolving downstream genes in operons has been reported in other nematodes [14–19], but is not seen in other eukaryotic groups. For example, platyhelminth *Schistosoma mansoni* and the tunicates *Ciona intestinalis* and *Oikopleura dioica* each possess only a single SL, which is used to resolve polycistronic RNAs but is also added to monocistronic transcripts [20–22]. Similarly, the chaetognath *Spadella cephaloptera* and the cnidarian *Hydra vulgaris* splice a diverse set of SLs to both monocistronic and polycistronic transcripts [23, 24]. Remarkably, all protein-coding genes in trypanosomes are transcribed as polycistronic RNAs and resolved using a single SL, making SL *trans*-splicing an obligatory process for all mRNAs [25]. In contrast, dinoflagellates use SL *trans*-splicing for all nuclear mRNAs, but only a subset of genes are organised as polycistrons [26, 27]. Although SL *trans*-splicing also occurs in many other organisms including rotifers, copepods, amphipods, ctenophores, cryptomonads and hexactinellid sponges, operons and polycistronic RNAs have not been reported in these groups [7, 8, 28, 29].

All these examples illustrate a rich diversity in the SL *trans*-splicing machinery and its role in facilitating polycistronic gene expression and broader RNA processing. A major barrier in dissecting the evolutionary history of these phenomena is the difficulty in systematically quantifying SL *trans*-splicing events. Identifying the full SL repertoire would traditionally require laborious low-throughput cloning-based Sanger sequencing of the 5' ends of mRNAs [e.g., 16, 30]. High-throughput RNA-Seq data is an attractive alternative resource that may often already exist for the focal organism. Some studies have demonstrated that SLs can, in principle, be identified from overrepresented 5' tails extracted directly from RNA-Seq reads [31, 32]. The recent SLFinder pipeline uses overrepresented k-mers at transcript ends as guides ("hooks") for annotating potential SL genes in genome assemblies [33]. SLFinder can detect known SLs

in several model organisms, but since it does not take into account the known functionally important sequence features of SL RNAs, its outputs may be noisy, incomplete and swamped by pseudogenes [33].

Once an SL sequence repertoire has been established, the next steps are to quantify SL *trans*-splicing events genome-wide and to establish functional links between these events and operonic gene organisation. Several studies have demonstrated that 5' information from RNA-Seq reads can be exploited to quantify SL *trans*-splicing events [28, 34], and the SL-QUANT pipeline has automated this task for *C. elegans* and other nematodes [35]. Similarly, it has been demonstrated in the nematodes *Pristionchus pacificus* and *Trichinella spiralis* that genome-wide patterns of SL *trans*-splicing events can be exploited to predict novel operons from SL splicing ratios [18, 19]. However, no software exists to implement these prediction strategies and render them universally applicable beyond the Nematoda.

Here we present two fully-automated pipelines that address all these shortcomings and present a unified and universal approach to examining SL *trans*-splicing and operonic gene organisation from RNA-Seq data in any eukaryotic organism. First, SLIDR is a more efficient, sensitive and specific alternative to SLFinder, implementing fully customisable and scalable *de novo* discovery of SLs and associated SL RNA genes. Second, SLOPPR implements a generalised and more flexible solution to quantifying genome-wide SL *trans*-splicing events than SL-QUANT. Uniquely, it provides algorithms for inference of SL sub-functionalisation and customisable prediction of operonic gene organisation. Both pipelines can process single-end or paired-end data from multiple libraries that may differ in strandedness and read configuration, thus allowing for flexible high-throughput processing of large RNA-Seq or EST datasets from multiple sources. These pipelines present a complete one-stop solution for systematically investigating SL *trans*-splicing and operon organisation in all eukaryotes.

Implementation

SLIDR: Spliced leader identification from RNA-Seq

SLIDR extracts evidence of SLs directly from RNA-Seq reads that contain unmapped 5' tails after alignment to a genome or transcriptome reference (broadly similar to 32). Unlike other methods, SLIDR then implements several optional plausibility checks based on functional nucleotide motifs in the SL RNA molecule, i.e., splice donor and acceptor sites, *Sm* binding motifs and a number of stem loops. These features are expected to be present due to shared evolutionary ancestry of SL RNAs with the snRNAs involved in intron removal by *cis*-splicing [6, 36]. For each plausible SL sequence, expressed SL RNA genes and SL *trans*-spliced genes are annotated in the reference, providing a means of manual inspection of the SL *trans*-splicing landscape if desired.

RNA-Seq reads are aligned to the genome or transcriptome reference using HISAT2 [37] or BOWTIE2 [38] in local alignment mode. Since soft-clipped read tails must be long enough to capture full-length SLs (typically about 22 bp in nematodes), a relaxed alignment scoring function is implemented that allows for up to 25 bp tails in a 75 bp read and can easily be customised by the user to accommodate more extreme SL lengths, for example 16 bp in *Ciona intestinalis* [39] or 46 bp in *Hydra vulgaris* [24]. Tails from the read end corresponding to the 5' end of the transcript (inferred from library strandedness) are extracted using SAMTOOLS [40] and dereplicated, 3'-aligned and clustered at 100% sequence identity using VSEARCH [41]. Each cluster thus represents a single putative SL, comprising a collection of 3'-identical read tails of varying length that only differ in their 5' extent (Figure 1).

The cluster centroids (longest sequence in each cluster) are then subjected to a number of functional plausibility checks. The centroids are aligned to the genome or transcriptome reference using BLASTN

[42] with 100% sequence identity and a relaxed customisable E-value (e.g., 1) to accommodate short queries. Matches that contain the full 3' end of the centroid are retained and the putative full SL RNA sequence (of customisable length) is extracted from the reference using BEDTOOLS [43]. The SL RNA sequence is then inspected for customisable splice-donor (e.g., GT) and *Sm* binding (e.g., AT{4,6}G) sites [44, 45], and secondary structure stem loops are predicted using RNAFOLD [46]. Default criteria expect the *Sm* motif c. 50 bp downstream of the splice donor site, and one stem loop on each side of the *Sm* binding motif [30, 45, 47]. In the reference sequence immediately upstream of the aligned portion of each RNA-Seq read, a splice acceptor site (e.g., AG) is required, corresponding to the *trans*-splice acceptor site of the gene (Figure 1).

The locations of splice donor and splice acceptor sites may not be as expected if the 3' end of the SL and the 3' end of the *trans*-splice acceptor site happen to be identical. In these cases, the RNA-Seq read alignment overextends in 5' direction into the *trans*-splice acceptor site and thus 3' truncates the soft-clipped SL read tail (Figure 1). These missing 3' nucleotides can be reconstructed from surplus nucleotides located between the 3' end of the tail BLASTN match and the splice donor site, and must be identical to those surplus nucleotides located between the 5' read alignment location and the splice acceptor site (Figure 1). Following plausibility checks and 3' reconstruction where necessary, all tail cluster centroids are subjected to another round of 3' alignment and clustering at 100% sequence identity in VSEARCH before final SL consensus construction is carried out in R [48]. Final SLs must be supported by at least two reads and must be spliced to at least two genes that are not located in the immediate vicinity (1 kbp distance) of the SL RNA gene [31].

SLOPPR: Spliced leader-informed operon prediction from RNA-Seq

SLOPPR is designed as a genome-annotation tool that predicts operons from genome-wide distributions of SL *trans*-splicing events at pre-annotated genes. RNA-Seq reads that contain evidence of 5' SLs are identified using a sequence-matching approach equivalent to the "sensitive" mode of SL-QUANT [35]. The operon prediction algorithm is built upon the SL1/SL2-type functional specialisation of SLs observed in many nematodes, but is fully customisable to accommodate other relationships between SLs and operonic genes, even when SL specialisation is absent. Unlike previous approaches that have defined operons in various organisms primarily via short intercistronic distances [17, 18, 21, 49], SLOPPR defines operons principally via SL *trans*-splicing patterns and only optionally takes intercistronic distance into account. SLOPPR can also identify and correct gene annotations where operonic genes are incorrectly annotated as a fused single gene (cf., 19), paving the way for *trans*-splicing-aware genome (re-)annotation.

RNA-Seq reads containing SLs are identified using a three-tier strategy. Since such reads cannot align end-to-end to the genome because of the *trans*-spliced 5' SL tail, all reads are first aligned end-to-end to the genome reference using HISAT2 [37] and unmapped reads are retained as candidates. If paired-end reads are used, the read corresponding to the 3' end of the transcript (inferred from library strandedness) must be aligned, and the read corresponding to the 5' end of the transcript must be unaligned [35]. The 5' ends of the unaligned candidate reads are then screened for overlap with the 3' portion of any number of supplied SL sequences using CUTADAPT [50]. Finally, those reads that align to the genome end-to-end after the SL tail has been trimmed are then quantified against exons and summarised at the gene level using FEATURECOUNTS [51]. Likewise, background expression levels of all genes are obtained from the original end-to-end read alignments and from candidate reads without SL evidence. This screening strategy is carried out for each RNA-Seq library independently, thus allowing for comparisons among biological replicates during analysis (Figure 2A).

The nature of the SL *trans*-splicing process means that SLs must only be present at the first exon of

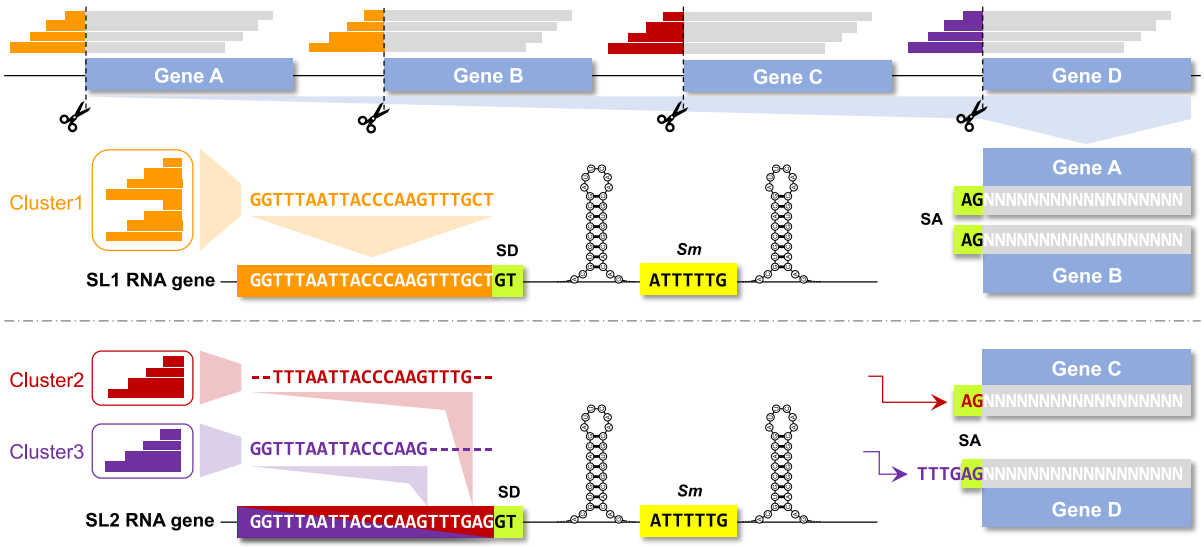

Figure 1: Schematic representation of the SLIDR pipeline (Spliced leader identification from RNA-Seq). Local alignments of reads (grey) to a genomic reference (illustrated by four genes A-D) allow for 5' SL tails to be soft-clipped and extracted (coloured read portions). Similarity clustering of 3' aligned read tails from all genes produces unique consensus SL candidates (cluster centroids), which are required to align to the genomic reference to identify candidate SL RNA genes (illustrated by SL1 and SL2 genes). In SL RNA genes, a splice donor site (SD; for example GT) is expected immediately downstream of the genomic alignment, followed by an *Sm* binding site (for example 5'-ATTTTTG-3') bookended by inverted repeats capable of forming stem loops in the RNA transcript. Conversely, the spliced gene requires a splice acceptor site (SA; for example AG) immediately upstream of the 5' read alignment location in the genomic reference. In this illustration, the example SL1 is fully reconstructed from a single read-tail cluster (cluster 1) with GT and AG splice sites in the expected locations (genes A and B). In contrast, the example SL2 highlights how read tails may be 3'-truncated due to overlap with the splice acceptor site (genes C and D) and the upstream *trans*-splice acceptor site sequence at some genes (gene D). These missing nucleotides can be filled in from the *trans*-splice acceptor site region guided by the distance between the 3' tail alignment location and the splice donor site (GT). Note that although cluster 2 is also 5' truncated due to insufficient coverage at gene C, consensus calling with cluster 3 allowed for reconstructing the full SL2 RNA gene.

a gene, i.e. the 5' end. Incorrect gene annotations thus become obvious when internal exons receive SL reads. SLOPPR implements an optional gene-correction algorithm that splits gene annotations at exons with distinct SL peaks compared to neighbouring exons (Figure 2B). To obtain exon-based SL counts, genome annotations are converted to GTF using GFFREAD from CUFFLINKS [52], unique exons are extracted using BEDTOOLS [43] and SL reads are quantified with FEATURECOUNTS at the exon level instead of gene level. The peak-finding algorithm is designed to correctly handle reads that may span multiple exons (Figure 2B).

The SL read counts obtained from FEATURECOUNTS are normalised against library size using CPM (counts-per-million) against the background gene counts [53]. The normalised SL read-count matrix is then subjected to generalized principal component analysis (PCA) and hierarchical clustering designed for sparse count matrices [54], treating SL read sets as samples and genes as variables. This summary of genome-wide distributions of SL *trans*-splicing events allows for identifying the distinct *trans*-splicing patterns of SL2-type SLs expected from their specialisation to resolve downstream operonic genes. If SL2-type SLs are not known, K-means clustering and linear discriminant analysis are used to assign SLs to one of two synthetic clusters assumed to correspond to SL1-type and SL2-type SLs (Figure 2C). Visual inspection of the clustering results allows the user to determine consistency across biological replicates (if available) and to ascertain functional groups of SLs beyond the SL1/SL2-type groups.

Based on the SL clustering results and pre-defined SL1/SL2-type groups (if known), the SL2:SL1 CPM ratio is computed and summarised across all genes that receive both SL types. The operon prediction algorithm is based on finding uninterrupted runs of adjacent genes with SL2-bias, which are designated as downstream operonic genes (Figure 2D). By default, no SL1-type reads at all are allowed, but a more relaxed SL2:SL1 ratio cutoff can be provided. The optimal cutoff is species-specific and could be identified empirically from inspecting the distribution of SL2:SL1 read ratios or from observed read ratios at known operonic genes [19]. After tracts of SL2-biased downstream operonic genes have been designated, each tract can, optionally, receive an additional upstream operonic gene that shows SL1-type bias or absence of SL *trans*-splicing (Figure 2D).

Finally, intercistronic distances among the predicted operonic genes are computed and compared to genome-wide intergenic distances to diagnose tight physical clustering of operonic genes (Figure 2E). These distances are obtained from the boundaries of consecutive "gene" GFF annotation entries, so their accuracy depends entirely on the provided genome annotations, which should ideally define gene boundaries by poly(A) and *trans*-splice acceptor sites. If desired, operon prediction can take intercistronic distances into account, either via a user-supplied distance cutoff or via an automatic K-means clustering method that splits the genome-wide distribution of intercistronic distances into two groups, corresponding to tight gene clusters (potential operons) and non-operonic genes. As such, by manually specifying SL1/SL2-type SLs, SL2:SL1 ratio cutoff, upstream gene designation and intercistronic distance cutoff, a large gamut of relationships between SLs and operonic genes can be explored, even in situations where no subfunctionalisation of SLs for operon resolution exists, for example in kinetoplastids or tunicates [20–22].

Results and Discussion

Validation of SLIDR in nematodes

In order to assess the performance of SLIDR in identifying SL RNAs, we validated and benchmarked the pipeline in several nematodes where reference genome assemblies are available and the SL repertoire is

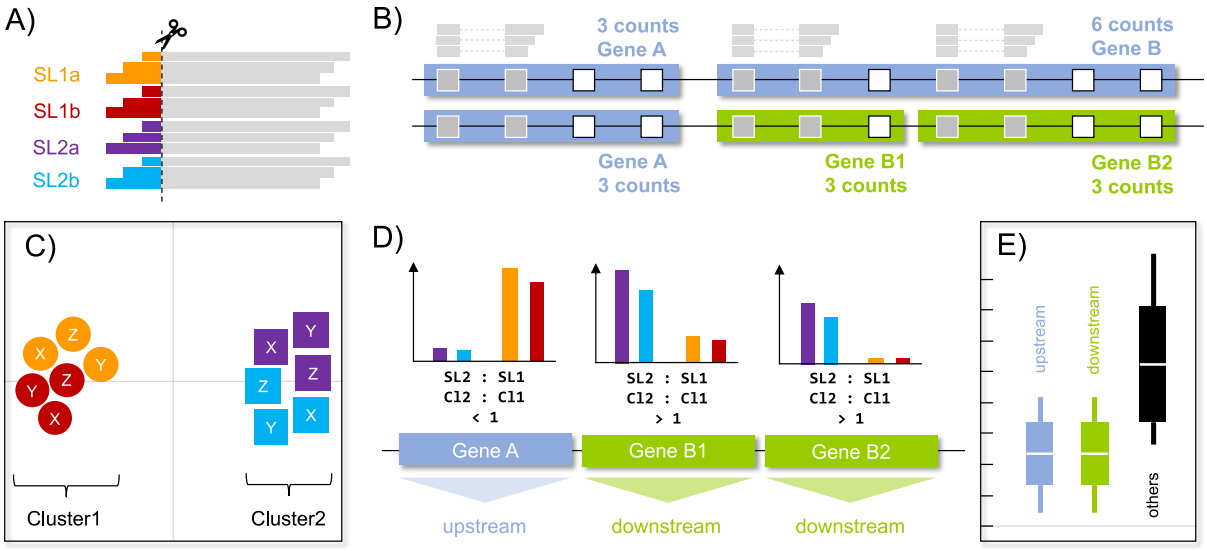

Figure 2: Schematic representation of the SLOPPR pipeline (Spliced leader-informed operon prediction from RNA-Seq). A) Spliced leader tails (example: SL1a, SL1b, SL2a and SL2b) are identified and trimmed from the 5' end of reads that correspond to the 5' end of transcripts. B) Trimmed reads are aligned to the genome, quantified against exons (squares; grey: covered; white: not covered) and counts are summarised by gene (example: two genes A and B). Incorrect gene annotations (fused operonic genes) can optionally be identified and corrected via SL reads at internal exons (example: Gene B is split into B1 and B2). C) SL read sets from multiple libraries (example: X, Y and Z) are ordinated via PCA on genome-wide read counts and grouped into two clusters (K-means clustering) expected to correspond to SL1 (circles) and SL2-type (squares) subfunctionalisation. D) SL2:SL1 read ratios are computed between pre-defined SL groups (SL1, SL2) or inferred clusters (Cl1, Cl2). Operons are predicted via tracts of genes receiving SL2 bias (downstream operonic genes) plus an optional upstream gene receiving either an SL1 bias or no SLs at all. E) Intercistronic distances among predicted operons are expected to be reduced compared to intergenic distances among non-operonic genes (others). Operon predictions can optionally be filtered by intercistronic distance using a user-supplied or inferred optimal cutoff.

well characterised: *Caenorhabditis elegans, Caenorhabditis briggsae, Pristionchus pacificus, Meloidogyne hapla*, *Trichinella spiralis* and *Trichuris muris*. In all cases, we demonstrate that SLIDR detects all known SLs and often discovers novel SL variants (Table 1, Supplemental Table S1). We also provide a proof-of-principle of the transcriptome-mode of SLIDR in *C. elegans*, where the transcriptome is well resolved, and *Prionchulus punctatus*, where no curated reference transcriptome exists. All RNA-Seq data were retrieved from public sources (NCBI SRA or ENA). Illumina adapters (5'-AGATCGGAAGAGC-3') and poor-quality bases (phred 20) were trimmed from all datasets with TRIM_GALORE 0.6.4 [55]. All SLIDR runs used the default splice donor (GT) and splice acceptor (AG) sites and default parameters unless otherwise specified.

Caenorhabditis elegans

C. elegans possesses the best understood repertoire of nematode SLs, comprising two types, SL1 and SL2, both of which are encoded by multi-copy gene families with well-described SL RNA structures and *Sm* binding motifs [11–13, 53]. SL *trans*-splicing affects up to 84 % of genes [34, 53], which makes *C. elegans* SLs an appropriate benchmark repertoire that any SL detection pipeline should fully resolve.

We downloaded the genome assembly GCF_000224145.3 and six out of 24 unstranded 2x100 bp datasets from bioproject PRJNA270896 (SRR1727796–SRR1727801). These are the same datasets used by Calvelo et al. [33] to benchmark their SLFinder pipeline. We ran SLIDR with parameter -S '.{40,55}AC?T{4,6}G' and compared the identified SL RNA genes with the reference gene annotations using BEDTOOLS INTERSECT 2.28.0 [43]. SLIDR identified 5,447 reads that were assembled into a single full-length SL1 sequence and complete 'donatrons' (the intron-like portions of the SL RNAs) were detected for all 10 functional *sls-1* genes (*sls-1.1* and *sls-1.5* are incorrectly annotated 5'-truncated pseudogenes). SLIDR also detected all 11 SL2 sequence variants corresponding to 18 out of 19 annotated *sls-2* genes from as few as eleven reads (Supplemental Table S1). Two reads aligned to the pseudogene *sls-2.19* but dropped out because no splice acceptor sites were present at their gene targets. Thus, SLIDR detected all 18 functional *C. elegans* SL2 copies and correctly omitted pseudogenes from its output (Supplemental Table S1).

These results are clearer than those reported by SLfinder using the same data [33]. Although SLFinder detected the full SL1 sequence in the transcriptome data and the same *sls-1* genes in the genome as SLIDR, the transcript tails ("hooks") were noisy, did not allow for identifying splice donor sites and required some manual curation [33]. Most strikingly, SLFinder only detected 5 SL2 sequence variants and only 8 out of 19 *sls-2* genes [33]. SLFinder did detect splice donor sites in most of these cases, though the sites were often overlapped by the hook sequences [33]. Thus, SLIDR not only detected all *C. elegans* SL sequences and functional genes with superior sensitivity and specificity from very few reads, but also obviated the requirement by SLFinder to assemble *de novo* transcriptomes at high computational cost (six assemblies were required for the SLFinder analysis).

Finally, we tested SLIDR with a transcriptome reference instead of a genome reference. In this situation, SLIDR cannot confirm splice acceptor sites because non-coding sequence regions are not expected to be present in a transcriptome reference; however, if the reference happens to contain SL RNAs it will still be possible to find splice donor and *Sm* binding sites. We used the curated transcriptome GCF_000002985.6_WBcel235 (contains *sls-1* and *sls-2* RNAs) and and three stranded 2x150 bp RNA-Seq libraries from bioproject PRJEB28364 (ERR2756688, ERR2756689, ERR2756736) from NCBI to provide the best possible input data. SLIDR was run with the same *Sm* motif regular expression as above and was able to detect one SL1 variant (*sls-1.10* RNA) and nine SL2 variants from all 18 functional *sls-2* RNAs. These results are comparable to those obtained with a genome reference but the inability to confirm splice-acceptor sites means that a large number of false positive candidate SLs were

Table 1: Spliced leader identification (SLIDR) results in seven nematodes and three other eukaryotes. Identifiers for reference genomes/transcriptomes and RNA-Seq libraries are presented alongside numbers of quality-trimmed reads (QC), the *Sm* motif regular expression notation used to filter SLs, numbers of expected SLs detected, numbers of novel SLs identified and numbers of expected SL RNA genes detected. Question marks represent unknown or poorly characterised SL RNA gene numbers.

Species	Reference	Bioproject	QC reads	*Sm* motif regex	SLs (detected | expected)	novel SLs	SL RNA genes (detected | expected)
Caenorhabditis elegans	GCF_000224145.3	PRJNA270896	150,041,952	.{40,55}AC?T{4,6}G	12 | 12	-	28 | 28
Caenorhabditis elegans	GCF_000002985.6[1]	PRJEB28364	45,443,137	.{40,55}AC?T{4,6}G	10 | 12	-	19 | 28
Caenorhabditis briggsae	PRJNA10731.WBPS5	PRJNA104933	21,265,790	.{40,55}AT{4,6}G	2 | 7	1	28 | 18+?
Caenorhabditis briggsae	PRJNA10731.WBPS5	PRJNA489172	78,497,095	.{40,55}AT{4,6}G	6 | 7	2	28 | 18+?
Caenorhabditis briggsae	PRJNA10731.WBPS5	PRJNA231838	59,271,467	.{40,55}AT{4,6}G	7 | 7	-	28 | 18+?
Pristionchus pacificus	Hybrid1	SRP039388	81,387,873	.{40,55}[AG]T{4,6}[AG]	6 | 11	6	219 | 203
Pristionchus pacificus	GCA_000180635.3	PRJNA338247	330,856,071	.{40,55}[AG]T{4,6}[AG]	6 | 11	6	640 | 203
Meloidogyne hapla	PRJNA29083.WBPS14	PRJNA229407	169,404,394	.{30,80}AT{4,6}G	5 | 5	9	15 | ?
Meloidogyne hapla	PRJNA29083.WBPS14	PRJEB14142	62,113,573	.{30,80}AT{4,6}G	5 | 5	9	15 | ?
Trichinella spiralis	PRJNA12603.WBPS10	PRJNA510020	201,164,867	.{20,50}AT{4,6}G	15 | 15	-	21 | 48
Trichuris muris	PRJEB126.WBPS15	PRJEB1054	115,460,947	.{25,50}AT{4,6}G	13 | 13	6	20 | 13
Prionchulus punctatus	de novo Trinity[1]	PRJEB7585	71,087,651	.{25,60}AC?T{4,6}G	2 | 6	-	2 | ?
Ciona intestinalis	GCF_000224145.3_KH	PRJNA396771	108,760,969	.{2,25}AGCTTTGG	1 | 1	-	11 | 15
Ciona intestinalis	GCF_000224145.3_KH	PRJNA396771	108,760,969	.{20,50}AT{4,6}G	1 | 1	3	2 | 15
Ciona intestinalis	GCF_000224145.3_KH	PRJNA297221	49,713,049	.{2,25}AGCTTTGG	1 | 1	-	11 | 15
Ciona intestinalis	GCF_000224145.3_KH	PRJNA433724	19,072,215	.{2,25}AGCTTTGG	1 | 1	-	11 | 15
Ciona intestinalis	GCF_000224145.3_KH	PRJNA376667	194,089,029	.{2,25}AGCTTTGG	1 | 1	2	14 | 15
Ciona intestinalis	GCF_000224145.3_KH	PRJNA376667	194,089,029	.{20,50}AT{4,6}G	1 | 1	1	3 | 15
Hydra vulgaris	Hm105	PRJNA497966	125,523,578	.{10,35}[AG]ATTTT[CG][AG]	2 | 12	1	4 | ?
Hydra vulgaris	Hm105	PRJNA641135	35,234,240	.{10,35}[AG]ATTTT[CG][AG]	3 | 12	3	5 | ?
Schistosoma mansoni	PRJEA36577.WBPS14	PRJNA225599	38,015,650	.{10,30}AGTTTTCTTTGG	1 | 1	-	112 | ?

[1] transcriptome reference

reported (103 candidate SLs in total; Supplemental Table S1). Nevertheless, the sensitivity of SLIDR even in transcriptome mode remains higher to that of SLFinder.

Caenorhabditis briggsae

C. briggsae is a close relative of *C. elegans that possesses a similar SL repertoire and shows considerable synteny of operons* [15, 17]. *Cbr*-SL1 is encoded by a repetitive gene cluster (about 65 copies; 56) linked to 5S rRNA genes [57], whereas SL2-type SLs are encoded by 18 genes and represent six distinct SL variants (*Cbr*-SL2, *Cbr*-SL3, *Cbr*-SL4, *Cbr*-SL10, *Cbr*-SL13, *Cbr*-SL14), of which four are shared with *C. elegans* [15]. Only 37 % of genes are SL *trans*-spliced [17], compared to 70–84 % in *C. elegans* [34, 53], though this is likely simply a reflection of differential transcriptome read depth.

We downloaded genome assembly PRJNA10731.WBPS5 from WormBase and the same two unstranded 2x42 bp libraries (SRR440557, SRR440441) from bioproject PRJNA104933 that Uyar et al. [17] used to identify genome-wide SL *trans*-splicing events. These data are particularly difficult to analyse because the short read length is likely to impede identification of the full-length SL. To maximise SL detection, SLIDR was run with parameter -x 2, which would allow a tail of at most 28 bp and leave at least 14 bp for read alignment. Irrespective, SLIDR only detected 5' truncated versions of *Cbr*-SL1 (45 reads spliced to 38 genes) and *Cbr*-SL3 (58 reads spliced to 55 genes).

To corroborate these findings with longer and higher-coverage read data, we ran SLIDR on two additional sets of libraries. First, using the high-coverage stranded 2x50 bp library SRR7781208 (bioproject PRJNA489172), SLIDR detected full-length *Cbr*-SL1 (251 reads spliced to 193 genes) and five out of six full-length SL2-type SLs (*Cbr*-SL13 was absent) supported by 8–908 reads and spliced to 8–475 genes (Supplemental Table S1). Second, we used an extensive set of five unstranded 2x75 bp and four stranded 2x125 libraries (bioprojects PRJNA231838 and PRJNA306868) in the same SLIDR run. These data supported full-length *Cbr*-SL1 (779 reads spliced to 310 genes) and all six full-length SL2-type SLs (6–72 reads spliced to 6–42 genes).

Overall, SLIDR detected known SLs with high sensitivity even from relatively unsuitable data in a species with relatively low SL *trans*-splicing frequency. We also note that SLIDR consistently detected five SL RNA gene loci for *Cbr*-SL1 and 23 instead of 18 loci for SL2-type SLs (two additional loci for *Cbr*-SL2 and three additional loci for *Cbr*-SL3; 17, 57).

Pristionchus pacificus

P. pacificus possesses seven SL1-type (*Ppa*-SL1) and four SL2-type (*Ppa*-SL2) SLs, which are encoded by 187 and 16 gene loci respectively [18]. Using *Ppa*-SL1a and *Ppa*-SL2a-enriched RNA-Seq, Sinha et al. [18] showed that 90 % of genes are SL *trans*-spliced. Since it is unknown to what extent non-enriched RNA-Seq data may underestimate this proportion, we took the opportunity to compare SLIDR between enriched and non-enriched RNA-Seq data. We downloaded the Hybrid1 genome assembly [58] and SNAP annotations from http://www.pristionchus.org, and the three enriched and non-enriched unstranded 2x75 bp libraries from bioproject SRP039388 [18]. SLIDR was run with parameter -S '.{40,55}[AG]T{4,6}[AG]' to capture both SL1 and SL2 *Sm* binding motifs [14].

Using the SL1-enriched library alone, SLIDR detected *Ppa*-SL1a (1,184 reads spliced to 143 genes), *Ppa*-SL1b (42 reads spliced to 24 genes) and three novel SL1 variants (Supplemental Table S1). Despite the SL1 enrichment, there was also evidence of *Ppa*-SL2b/m and one novel SL2 variant, together representing 111 reads spliced to 46 genes. Contamination was also obvious for the SL2a-enriched library, where the

most frequent SL was *Ppa*-SL1a (336 reads spliced to 69 genes) followed by *Ppa*-SL2a (112 reads spliced to 44 genes). SLIDR also detected four novel SL1 variants (three of which were also detected in the SL1-enriched library) and *Ppa*-SL2b/c. The non-enriched library produced very similar results at similar read and gene depths, comprising *Ppa*-SL1a, *Ppa*-SL2a/c and two novel SL1 and SL2 variants each (Supplemental Table S1). Overall, 190 SL1 RNA genes and at least 29 SL2 RNA genes were detected, which is a slight increase compared to those reported by Sinha et al. [18].

These results highlight that the SL enrichment via biotin pulldown may not have worked quite as effectively as suggested by qPCR control experiments [18]. Although it is impossible to quantify the degree of contamination with non-*trans*-spliced transcripts, the SLIDR results suggest that the 90% SL *trans*-splicing rate may be an overestimate. Using these three low-coverage libraries, SLIDR estimates the SL *trans*-splicing rate to only about 1%. To explore how SLIDR performs with higher coverage data, we then ran SLIDR with six unstranded 2x150 bp libraries from bioproject PRJNA338247 (SRR4017216–SRR4017221) and a better resolved genome assembly plus annotations (GCA_000180635.3). SLIDR detected the same breadth of known and novel SL1 and SL2 variants as above, but recovered considerably more reads and SL *trans*-spliced genes (e.g., 16,085 *Ppa*-SL1a reads spliced to 5,148 genes). Up to 6,498 genes are SL *trans*-spliced in these libaries, resulting in an estimated rate of c. 25 %, which is still substantially below the postulated rate of 90 % [18]. Due to the superior genome assembly, SLIDR detected at least 611 SL1 RNA genes and at least 29 SL2 RNA genes (Supplemental Table S1).

Overall, SLIDR detected a rich diversity of known and previously unreported SL1-type and SL2-type SLs beyond the canonical *Ppa*-SL1a and *Ppa*-SL2a variants [15, 18]. SLIDR suggested that five out of the seven *Ppa*-SL1 loci predicted by Sinha et al. [18] are not expressed. While likely not exhaustive, these results highlight the sensitivity of SLIDR in detecting functional SL variants even from low coverage data.

Meloidogyne hapla

The plant-root knot nematode *M. hapla* possesses the canonical *C. elegans* SL1 and four additional variants, all of which are *trans*-spliced to a minority of only 10 % of genes [31]. We used genome PRJNA29083.WBPS14 and the same 32 SRA runs from bioproject PRJNA229407 that were used to discover these SLs [31]. These data are particularly difficult to analyse since they are 75 bp single-ended, unstranded and originate from mixed-culture RNA samples containing primarily material from the host plant *Medicago truncatula*. Since reads from unstranded single-end libraries originate from the 5' end of the transcript only 50% of the time, usable coverage is effectively halved. SLIDR was run with parameters -S '.{30,80}AT{4,6}G' -R 90 to allow for larger variation in *Sm* binding motif location and longer SL RNA.

SLIDR detected all five known SLs and discovered at least nine novel SLs, suggesting that the SL repertoire in this organism is much larger than previously identified (Supplemental Table S1). However, only at most 176 reads were detected per SL, and at most 143 genes were SL *trans*-spliced. This is consistent with low incidence of *trans*-splicing in this organism [31] and is not due to the RNA-Seq data. We confirmed this with longer reads (100 bp single-end) from a different bioproject (PRJEB14142; biosamples SAMEA4003664 and SAMEA4003666): SLIDR detected the same known and novel SLs with even fewer reads (at most 74) and fewer *trans*-spliced genes (at most 68), suggesting that very high coverage datasets would be necessary to exhaustively characterise SL *trans*-splicing events in this organism (Supplemental Table S1). Nevertheless, SLIDR detected known and novel SLs even at this low coverage, illustrating high sensitivity even with poor data.

Trichinella spiralis

The parasite *T. spiralis* possesses a diverse and unusual set of 15 SLs that are encoded by up to 48 genes [30] and are spliced to c. 30 % of all genes [19]. Three out of these 15 SLs (*Tsp*-SL2, *Tsp*-SL10 and *Tsp*-SL12) are SL2-type SLs specialised for resolving downstream genes in operons [19]. We downloaded genome PRJNA12603.WBPS10 and three RNA-Seq libraries (SRR8327925-SRR8327927) from bioproject PRJNA510020 [19]. SLIDR was run with parameter -S '.{20,50}AT{4,6}G' to accommodate for the smaller distance of the *Sm* binding motif to the splice donor site [30]. SLIDR detected all 15 known SLs and a total of 21 SL RNA genes (Supplemental Table S1), which is an increase over the original 19 SL RNA genes identified from cDNA evidence [30]. These numbers also suggest that many of the 29 additional genomic loci predicted by Pettitt et al. [30] may not be functional. We note relatively low numbers of SL reads (12-516) and SL *trans*-spliced genes (14-387), which is consistent with the notion that SL *trans*-splicing affects at most about 30 % of genes in this organism (Supplemental Table S1).

Trichuris muris

T. muris is a gastrointestinal parasite closely related to *Trichinella spiralis* and possesses 13 SLs that, unlike those of *T. spiralis*, resemble *C. elegans* SLs and are encoded by 13 genes [59, 60]. Three of these SLs (*Tmu*-SL1, *Tmu*-SL4 and *Tmu*-SL12) are SL2-type SLs [59]. The genome-wide extent of SL *trans*-splicing in this organism is unknown. We downloaded genome assembly PRJEB126.WBPS15 from WormBase and five unstranded 2x100 bp libraries from bioproject PRJEB1054. SLIDR was run with -S '.{25,50}AT{4,6}G' to account for a shorter distance of the *Sm* binding motif to the splice donor site [59]. SLIDR detected all 13 known SLs from 9–301 reads spliced to 10–249 genes (Supplemental Table S1). Additionally, at least six novel SLs were identified from 4–1,117 reads spliced to 12–805 genes. The numbers of SL RNA genes ranged between 1 and 3, suggesting that some of the SLs are encoded by multi-copy genes. Overall, more than 2,000 genes received SLs, which would suggest an SL *trans*-splicing rate of about 15 % (Supplemental Table S1).

Prionchulus punctatus

A limited SL repertoire of *P. punctatus* has been determined using 5-RACE of cDNA and comprises six SLs that show structural similarity with *C. elegans* SL2 [16]. However, since no genome assembly exists, the genomic organisation of SL genes and the extent of SL *trans*-splicing are unknown [16]. Only two RNA-Seq libraries are available (SRA accessions ERR660626, ERR660627, bioproject PRJEB7585) and no reference transcriptome assembly exists. We tested the performance of SLIDR with these two libraries (2x100 bp) using a *de novo* transcriptome assembly obtained from the same libraries. Illumina adapters and poor-quality bases (phred 30) were trimmed using TRIM_GALORE 0.6.4 [55], transcripts were assembled using TRINITY 2.8.5 [61] and clustered at 100 % sequence similarity using CDHIT 4.8.1 [62]. The final assembly comprised 141,825 transcripts with an N50 of 786 bp (184-16,745 bp) and total transcriptome size of 74.31 Mbp.

We ran SLIDR with a relaxed *Sm* location range (-S '.{25,60}AC?T{4,6}G') but only discovered two (*Ppu*-SL1 and *Ppu*-SL3) out of six known SLs, supported by only 96/16 reads and spliced to 29/7 genes respectively (Supplemental Table S1). While these results are little more than initial proof-of-concept, it must be noted that the success of this *de novo* strategy depends critically on the presence of SL RNA sequences in the transcriptome data. Since SL RNAs are not polyadenylated, RNA-Seq library preparation protocols that rely on poly(A) selection will not capture SL RNAs, which limits the use of publically available datasets that were not generated with ribosomal depletion protocols [63,

64] or poly(A)-tailing prior to library preparation [16]. Thus, we expect SLIDR to underperform in transcriptome mode unless a high-quality transcriptome and high-coverage RNA-Seq data are available.

Validation of SLIDR in other eukaryotes

Although we designed the algorithms in SLIDR on the basis of known SL RNA structure in nematodes, most of the filters based on sequence motifs can be fully customised or even disabled to relax stringency if required. Here we demonstrate that SLIDR performs equally as well in other eukaryotes with SL repertoires and SL RNA structures that are divergent from the nematode consensus. We used the same datasets for *Ciona intestinalis, Hydra vulgaris* and *Schistosoma mansoni* that Calvelo et al. [33] used to benchmark their SLFinder pipeline in order to carry out a detailed comparison of the two pipelines. For all three species, SLIDR produced much clearer results with more sensitivity and specificity than SLFinder, similar to what we observed above in *C. elegans*.

Ciona intestinalis

The tunicate *C. intestinalis* possesses a single 16 bp spliced leader 5'-ATTCTATTTGAATAAG-3' that is spliced to at least 58% of expressed genes [21, 39, 65]. The SL RNA is very short (46 bp) and contains the *Sm*-binding motif 5'-AGCUUUGG-3' [66]. The SL RNA has been suggested to be encoded by a highly repetitive gene family comprising at least 670 copies, though the reference genome contains at most 15 of them due to assembly constraints [67]. SLFinder detected this single SL after extensive parameter tweaking and found two distinct gene variants comprising 14 loci with a splice donor site [33].

We downloaded the same genome assembly (GCF_000224145.3) and the same three out of six 100 bp paired-end datasets from bioproject PRJNA396771 (SRR5888437, SRR5888438 and SRR5888439) from NCBI as Calvelo et al. [33]. We noticed after preliminary read alignments to the genome that the libraries are not reverse stranded as described in the SRA entries, but are, in fact, unstranded. SLIDR was run with the parameters -x 0.6 -e 5 -O 5 -R 30 -S '.{2,25}AGCTTTGG' to enforce shorter soft-clipping (maximum 24 bp given 100 bp reads), a BLAST e-value cut-off of 5 (to allow short matches of c. 11 bp), maximum 5 bp outron overlap, 30 bp RNA length excluding the SL, and the *Sm*-like motif located up to 25 bp downstream of the GT splice donor site.

SLIDR identified a single SL from only 15 reads (spliced to 15 genes) despite very high genome alignment rates of 93-95 %; this SL represents the known SL sequence with some evidence of extra 5' nucleotides (5'-taaggcATTCTATTTGAATAAG-3'). All but one of the eleven loci identified by SLIDR were on chromosome NC_020175.2 (one was on NC_020166.2), and all loci were part of the 264 bp repeat unit that contains functional SL copies in the genome [67]. In contrast, the 14 loci detected by SLFinder were not located within the 264 bp repeat (none of the loci were on the correct chromosomes) and are therefore probably pseudogenes, which are rife in the genome [67]. Since the poor SL detection rate is at odds with the expected 58 % SL *trans*-splicing rate [65], we re-ran SLIDR with the remaining three libraries of the same bioproject (SRR5888437, SRR5888438 and SRR5888439), but found no improvement. Similarly, two libraries from two different bioprojects (SRR6706554, SRR2532443) yielded only slightly better results, detecting the same SL and SL RNA genes from 59 and 227 reads spliced to 54 and 132 genes respectively (Supplemental Table S1). Removing the filter for the *Sm* motif yielded the same SL sequence and detected >100 pseudogenes, but did not increase the number of SL *trans*-spliced genes (Supplemental Table S1). This difficulty in detecting SL *trans*-splicing would be compatible with the observation that expression levels of SL *trans*-spliced genes are 2-3x lower than those of genes that are not SL *trans*-spliced [39].

However, when we tried 13 libraries from bioproject PRJNA376667 we obtained substantially superior results: the known SL was detected from 38,467 reads spliced to 5,824 genes and originating from the same eleven loci as above. Additionally, two novel variants of this SL were detected from 262 reads spliced to 209 genes and originating from three novel loci, totalling 14 out of 15 SL RNA loci (Supplemental Table S1). Assuming 15,254 genes in the genome [21], these results indicate a SL *trans*-splicing rate of 40 %, which is much closer to the expected 58 % [65]. This substantial variability of SL *trans*-splicing rates between biosamples may be due to variability among life stages, tissues or RNA-Seq library preparation methods. We also noted that the 13 libraries had substantially lower genome alignment rates (30-70 %) than the other libraries, despite much greater evidence of SL *trans*-splicing, which could be due to contamination with material from organisms other than *C. intestinalis*.

When we omitted the *Sm* motif filter during exploration of the initial libraries (PRJNA396771), we made a curious discovery of potentially novel SLs that resemble nematode SLs instead of the canonical *C. intestinalis* SL. After increasing the SL RNA length (-R 60) and filtering for a nematode-like *Sm* motif (-S '.{20,50}AT{4,6}G') SLIDR detected one major 21 bp candidate (5'-CCGTTAAGTGTCTTGCCCAAG-3') defined by 2,068 reads and spliced to 6 genes, alongside two additional plausible candidates at much lower read depth (3–39 reads spliced to 4–6 genes) (Supplemental Table S1). That same major SL was also detected among the 13 libraries of bioproject PRJNA376667, defined by 897 reads and spliced to 20 genes (Supplemental Table S1). It was beyond the scope of this study to fully resolve and describe these novel SLs, but these preliminary results do highlight that SLIDR is much more sensitive than SLFinder, which found no evidence of these SLs in the same libraries. In summary, although both SLFinder and SLIDR detected the correct published SL sequence, SLFinder was unable to detect functional gene loci, whereas all loci detected by SLIDR are consistent with the known SL RNA properties for this species [66].

Hydra vulgaris

The cnidarian *H. vulgaris* possesses two types of spliced leaders that are added to at least one third of all genes: the first type (SL-A) is 24 bp long and is part of an 80 bp SL RNA [68], whereas the second type is much longer (46 bp SL, 107 bp SL RNA) and comprises a total of eleven SL variants across six SLs (SL-B to SL-G) [24]. The Sm binding sites differ between SL-A (5'-GAUUUUCGG-3') and all other SLs (5'-AAUUUUGA-3' or 5'-AAUUUUCG-3') [68]. SLFinder detected the full sequence of SL-B1 and at least 21 loci, all of which were 5' truncated and are thus probably pseudogenes. SLFinder found no evidence of SL-A or any of the other ten SLs [33].

We downloaded the genome assembly Hm105 and the same five stranded 2x100 bp datasets from bioproject PRJNA497966 (SRR8089745–SRR8089749) [33]. We ran SLIDR with the parameters -x 1.5 -R 60 -S '.{10,35}[AG]ATTTT[CG][AG]', which cover both Sm binding site motifs and should allow for detecting both the short and long SLs. SLIDR detected the full SL-B1 sequence from 799,327 reads spliced to 18,418 genes and identified two gene loci, both of which were also identified by SLFinder [33]. SLIDR also detected two 5' truncated versions of SL-D (10,696/4,494 reads, 2,727/1,677 spliced genes and two gene loci, none of which were identified by SLFinder) and a novel variant of SL-B1 (864 reads spliced to 224 genes; coded by a single locus that was also identified by SLFinder). These results highlight that SLIDR detected more SL variants that SLFinder and only reported functional SL RNA genes, whereas SLFinder identified a large number of truncated gene loci, which are likely to be pseudogenes.

Since the SL-B-type SLs are exceptionally long, one would require longer reads than 100 bp to detect full-length SLs with confidence. We tested another RNA-Seq library of 2x150 bp reads (SRR12070443) and were able to detect full-length SL-B1, SL-D and SL-E based on 5,446–261,507 reads spliced to

1,598–12,688 genes (Supplemental Table S1). We also detected a novel SL-A-type variant (1,916 reads spliced to 730 genes) and at least one novel SL-B-type variant (392 reads spliced to 202 genes). Similar to the libraries above, we observed strikingly large numbers of SL *trans*-spliced genes – by far the highest among any species detailed in this study. Contrary to previous estimates that only c. 33 % of c. 20,000 protein-coding genes are SL *trans*-spliced [24], our SLIDR results suggest that at least 63-92 % (12,688-18,418) genes may be SL *trans*-spliced. While these results indicate the need for further study, they demonstrate the level of inference possible with SLIDR.

Schistosoma mansoni

The platyhelminth *S. mansoni* possesses a single, relatively long (36bp) SL with an unusually long *Sm* binding site (5'-AGUUUUCUUUGG-3') and a total RNA length of 90 bp [69]. The transcripts from at least 46 % of genes undergo *trans*-splicing by this SL [22]. SLFinder detected this SL but missed the first two A nucleotides; similarly, all detected loci were considerably 5' truncated [33].

We downloaded the genome assembly PRJEA36577.WBPS14 and two unstranded 2x100 bp datasets from bioproject PRJNA225599 (SRR1020297–SRR1020298) [33]. SLIDR was run with parameters -x 1.25 -R 55 -S '.{10,30}AGTTTTCTTTGG' to allow for detecting this large SL and the unusual *Sm* binding site. SLIDR detected 18,568 reads that were assembled to the full length SL with two extra 5' nucleotides (Supplemental Table S1). This SL was encoded by 112 genes, all of which correspond to the Smansoni_pSL-1 gene cluster on chromosome SM_V7_6 detected by SLFinder [33]. Contrary to the 5' truncation that the SLFinder output suggested, these loci do contain the full length SL sequence, which was correctly identified by SLIDR. SLIDR also detected two sequence variants (from only 2-3 reads) of the SL corresponding to two loci, none of which were detected by SLFinder. SLFinder also reported 9 additional loci without clear splice donor sites, suggesting that these may be pseudogenes [33]. A total of 2,745 genes were SL *trans*-spliced, which is slightly higher than 2,459 genes previously identified from a large-scale RNA-Seq data (250 million reads; 22), but represents only 21 % instead of the expected 46 % SL *trans*-splicing rate [22].

Interestingly, SLFinder detected a genomic locus where the terminal ATG nucleotides of the SL sequence were replaced by ACG, though this was not informed by evidence from the RNA-Seq data [33]. This illustrates a key difference between the two software pipelines: SLFinder is primarily a genome annotation pipeline for SL loci that uses RNA-Seq evidence as initial anchors ("hooks") to search for all possible gene loci [33]. In contrast, SLIDR aims to extract SL evidence directly from RNA-Seq data and uses the genome only to extract additional evidence for functional sequence components of putative SLs; SLIDR therefore only annotates gene loci that are expressed in the RNA-Seq libraries under consideration and ignores alternative but unexpressed loci. Both approaches are clearly complementary, though our analyses suggest that SLIDR is more robust in detecting functional SLs and SL RNA genes.

Validation of SLOPPR in nematodes

We have previously used SLOPPR to comprehensively discover operons in the genome of a nematode, *T. spiralis*, for which there was only limited evidence for operon organisation [19]. Here we validate and benchmark SLOPPR in the nematodes *C. elegans, C. briggsae* and *P. pacificus*, all of which have well-characterised operon repertoires and use SL2-type *trans*-splicing to resolve mRNAs transcribed from these operons. SLOPPR correctly classified SL1- and SL2-type SLs in all three species and identified large proportions of known operons alongside several novel candidate operons (Table 2, Supplemental Table S2). We further validated SLOPPR by confirming the presence of SL2-type SLs in the nematode

T. muris, for which the genome-wide landscape of SL *trans*-splicing and operon organisation is unresolved [59].

Caenorhabditis elegans

C. elegans is the benchmark model organism for eukaryotic operons: up to 20 % of genes are situated in operons [53]. Downstream operonic genes are readily diagnosable by an 80%–95% bias toward SL2, though there are exceptions where downstream genes receive much lower proportions of SL2 [53].

We designed the operon prediction algorithm in SLOPPR on the basis of SL2:SL1 ratios at genes and benchmarked its performance with curated *C. elegans* operons. We downloaded the genome assembly and genome annotations from WormBase (PRJNA13758.WS276). These annotations contain 19,999 coding genes and 1,542 operons, which are absent from the NCBI version of the genome. We used a large dataset of 24 unstranded 2x100 bp RNA-Seq runs from bioproject PRJNA270896 (SRR1727796–SRR1727819). We provided SLOPPR with the canonical SL1 sequence and 11 SL2 variants as supplied by SL-QUANT [35], and with GFF annotations for the 1,542 reference operons. SLOPPR was run with the default SL2:SL1 ratio of infinity, thus enforcing absence of SL1 at downstream operonic genes.

SLOPPR classified 36 % of genes as strictly SL1 *trans*-spliced, only 1 % as strictly SL2 *trans*-spliced and 15 % as *trans*-spliced by both SL1 and SL2. The clustering algorithm correctly identified SL1- and SL2-type subfunctionalisation of SLs (Supplemental Table S2). From these classifications a total of 434 operonic genes were identified in 213 operons, with median intercistronic distance (distance between "gene" GFF annotations) of 105 bp. Of these operons, 166 (77 %) matched reference operons, but these represented only 11 % of the 1,524 total operons. This may be because the overall SL *trans*-splicing rate of 52 % was below the expectation of 70-84 %, and the proportion of genes receiving a mixture of SLs was much higher than the expected 6 % [53]. We thus relaxed the SL2:SL1 ratio threshold to 2, which predicted 721 operonic genes in 345 operons (99 bp median intercistronic distance), of which 295 (85 %) matched reference operons (19 % of 1,542 operons).

While these numbers illustrate that SLOPPR predicts bona fide operons and also finds novel candidate operons, they also demonstrate that this dataset is not nearly large enough to provide exhaustive insight into the SL *trans*-splicing landscape. Tourasse et al. [34] carried out a meta-analyis of SL *trans*-splicing in *C. elegans* using 1,682 RNA-Seq datasets comprising more than 50 billion reads, of which 287 million reads contained evidence of SLs. Even at this huge coverage 97.4% of SL *trans*-splicing events were supported by fewer than 100 reads and a vast number of events with very low read counts could not be distinguished from biological noise in the splicing process. This highlights the inherent limitations of standard RNA-Seq protocols and indicates that it is unrealistic to expect that all SL *trans*-splicing events and operonic genes be detected using limited amounts of RNA-Seq data.

Caenorhabditis briggsae

C. briggsae is an important comparative model to *C. elegans*, but its gene and operon repertoires are less resolved than those of its relative. The current genome annotations (PRJNA10731.WBPS14) contain only 48 confirmed dicistronic operons. In contrast, Uyar et al. [17] used tight gene clusters that receive SL2 to predict 1,034 operons, of which 51 % were syntenic with *C. elegans*. We decided to examine SLOPPR with both annotation sets. The current CB4 assembly and annotations were downloaded from WormBase (PRJNA10731.WBPS14); the CB3 assembly was downloaded vom UCSC to ensure that the annotations by Uyar et al. [17] were compatible.

Table 2: Operon prediction (SLOPPR) results in four nematode and two tunicate species. Identifiers for reference genomes and paired-end RNA-Seq libraries are presented alongside numbers of quality-trimmed reads (QC), numbers of reads with a spliced leader (SL), percentage of genes receiving an SL, numbers of available reference operons, numbers of predicted operons, specificity (predicted operons matching reference operons), sensitivity (fraction of reference operons detected), numbers of operonic genes, and median intercistronic distance among operonic genes. All runs did not require upstream operonic genes to be SL *trans*-spliced. Since the two tunicates *C. intestinalis* and *O. dioica* do not use SL2-type *trans*-splicing, operonic genes were automatically filtered by intercistronic distance (maximum 84 bp for *C. intestinalis* and 413 bp for *O. dioica*; see main text).

Species	Genome	Bioproject	QC reads	SL reads	SL genes	Reference operons	Predicted operons	Specificity | sensitivity	Operonic genes	Intercistronic distance (bp)
Caenorhabditis elegans	GCF_000224145.3	PRJNA270896	621,744,295	937,587	52 %	1,542	345	86 % | 19 %	721	99
Caenorhabditis briggsae	CB3 (UCSC)	PRJNA104933	21,265,790	209,005	27 %	1,035	840	81 % | 66 %	1,847	342
Caenorhabditis briggsae	CB3 (UCSC)	PRJNA231838	59,271,467	286,520	36 %	1,035	752	81 % | 59 %	1,631	332
Caenorhabditis briggsae	CB4	PRJNA104933	21,265,790	235,001	38 %	48	921	5 % | 96 %	2,058	111
Caenorhabditis briggsae	CB4	PRJNA231838	59,271,467	289,834	41 %	48	750	5 % | 81 %	1,626	110
Pristionchus pacificus	Hybrid1	SRP039388	81,387,873	438,557	20 %	2,219	190	61 % | 5 %	382	981
Pristionchus pacificus	Hybrid1	PRJNA338247	330,856,071	312,718	28 %	2,219	205	50 % | 5 %	414	872
Trichuris muris	PRJEB126.WBPS15	PRJEB1054	115,460,947	168,727	35 %	–	718	–	1,492	418
Ciona intestinalis	KH	PRJNA376667	194,089,029	452,505	51 %	1,328	1,172	94 % | 83 %	2,563	1
Oikopleura dioica	GCA_000209535.1	PRJNA269316, PRJDB5668	250,361,592	5,860,107	9 %	1,765	577	90 % | 30 %	1,292	33

We first used SLOPPR with the two unstranded 2x42 bp libraries from Uyar et al. [17]. SLOPPR quantified an overall SL *trans*-splicing rate of c. 27 % using the C3 assembly. SLOPPR predicted 1,346 operonic genes in 631 operons with a median intercistronic distance of 333 bp, of which 507 operons (80 %) matched the 1,035 reference operons (48 % detected). Relaxing the SL2:SL1 ratio threshold from infinity to two predicted 1,847 genes in 840 operons (342 bp median intercistronic distance), of which 682 (81 %) matched reference operons (65 % detected). Using the CB4 assembly, SLOPPR quantified an SL *trans*-splicing rate of 38 % and predicted 1,475 operonic genes in 688 operons (112 bp median intercistronic distance), of which 37 (5 %) were among the 48 (77 %) reference dicistronic operons. Relaxing the SL2:SL1 ratio to two resulted in 2,058 genes in 921 operons (111 bp median intercistronic distance) and recovered 46 out of 48 (95 %) reference dicistrons. These numbers highlight that the CB4 gene annotations are superior to those used by Uyar et al. [17] and that SLOPPR predicts *bona fide* operons and novel candidate operons (Supplemental Table S2).

However, two concerns were raised during analysis of these 2x42 bp RNA-Seq libraries: First, only 62-69 % of reads aligned to the genome and only 50-58 % were properly paired. This is consistent with the short read lengths and poor sequence quality which causes difficulty in aligning these reads with standard aligners [17]. Second, the SL *trans*-splicing patterns varied more between the two libraries (L1 vs. mixed life stages) than they did between SL1 and SL2-type SLs, which caused SLOPPR to cluster the SLs by library instead of SL type (Supplemental Table S2). We thus re-ran the analyses with longer reads from five unstranded 2x75 bp libraries (bioproject PRJNA231838). These libraries supported higher SL *trans*-splicing rates of 36 % and 41 % for the CB3 and CB4 genome assemblies respectively, and SLOPPR correctly identified SL1 and SL2-type clusters among these data using either genome assembly (Supplemental Table S2).

Using CB3, SLOPPR predicted 1,202-1,631 operonic genes in 564-752 operons (332-341 bp median intercistronic distance), of which 442-606 (78-80 %) were among the 1,035 CB3 reference operons (43-59 % detected). Using CB4, SLOPPR predicted 1,179-1,626 operonic genes in 553-750 operons (110 bp median intercistronic distance), of which 24-39 (4-5 %) were among the 48 reference dicistrons (50-81 % detected). These results are somewhat superior to those above, but echo essentially the same patterns. The more recent CB4 assembly clearly has better gene annotations that yield a much lower median intercistronic distance of about 110 bp, which is consistent with *C. elegans* [53]. However, this assembly contains only few curated operons and SLOPPR detected most of these. In addition, SLOPPR discovered a large set of novel operons, which warrants further study.

Pristionchus pacificus

P. pacificus is another important comparative model to *C. elegans* that resolves operons with SL2-type *trans*-splicing [15]. A comprehensive survey of SL *trans*-splicing events using SL1- and SL2-enriched RNA-Seq data suggested that 90 % of genes are SL *trans*-spliced and a total of 2,219 operons may exist on the basis of tight gene clusters and SL1/SL2 *trans*-splicing ratios [18]. We downloaded the Hybrid1 genome assembly [58] and SNAP genome annotations and operon annotations from http://www.pristionchus.org [18].

We first ran SLOPPR with those same SL-enriched unstranded 2x75 bp libraries from bioproject SRP039388, supplying the two canonical *Ppa*-SL1a and *Ppa*-SL2a sequences that were used for SL enrichment [18]. SLOPPR detected SLs at only 20 % of genes, even when including the non-enriched library (SRR1182510). The SL1-enriched library (SRR1542610) supported SL1 and SL2 *trans*-splicing at 16.17 % and 0.05 % of genes, consistent with SL1-enrichment. However, the SL2-enriched library (SRR1542630) showed no evidence of SL2-enrichment (0.98 % of genes) but comparable SL1 levels to the non-enriched control

library (8.2 % of genes). These results echo the SLIDR results using the same libraries (see above) and would suggest a far lower SL *trans*-splicing rate than 90 % [18]. Due to the low SL *trans*-splicing rate, only 234 operonic genes in 117 operons were predicted, of which 99 (84 %) matched the 6,909 operon-like gene clusters and 67 (57 %) matched the 2,219 plausible reference operons [18]. Relaxing the SL2:SL1 ratio threshold to two yielded slightly better results with 382 genes in 190 operons, of which 115 (61 %) were among the 2,219 reference operons.

To examine whether these poor numbers were due to the specific libraries, we then ran SLOPPR with a more extensive set of six unstranded 2x150 bp libraries from bioproject PRJNA338247 (SRR4017216–SRR4017221). Surprisingly, these libraries yielded similar results, suggesting that 28 % of genes are SL *trans*-spliced and predicting 300 operonic genes in 150 operons, of which 76 (50 %) matched the 2,219 reference operons. Using a relaxed SL2:SL1 ratio of two, 414 genes were identified in 205 operons, of which 102 (50 %) were among the 2,219 reference operons. Both sets of libraries yielded similar median intercistronic distances of 785-860 bp, which increased to 872-981 bp when relaxing the SL2:SL1 threshold. These distances are much larger than the 100 bp expected in *C. elegans* [53] but are consistent with the median distance of 1,149 bp among all 6,909 gene clusters in *P. pacificus* and very poor synteny of these clusters with *C. elegans* (only 37 out of 6,909 clusters are syntenic; 18). SLOPPR also correctly identified SL1- and SL2-type clusters from genome-wide *trans*-splicing patterns in both library sets, confirming that *Ppa*-SL1a and *Ppa*-SL2a are functionally diverged (Supplemental Table S2).

All these observations suggest that SLOPPR produces plausible results given the limitations of relying on RNA-Seq reads covering the 5' end of transcripts. The striking discrepancies between SLOPPR and the analyses by Sinha et al. [18] are likely due to their assumption that all reads from the SL-enriched libraries are from SL *trans*-spliced transcripts. Since only a small fraction of RNA-Seq reads originate from the 5' end of transcripts (confirmed by SLIDR and SLOPPR), this assumption cannot be confirmed bioinformatically, and thus it cannot be ruled out that these libraries contained substantial amounts of contaminant non-*trans*-spliced transcripts despite the authors' efforts of confirming their methods with qPCR [18]. If this were the case, their analyses based on SL1/SL2 *trans*-splicing patterns would be flawed, which would explain the poor overlap with the more transparent SLOPPR results. One way of testing this would be to combine SL-enrichment with long-read whole-transcript sequencing on the PacBio or ONT NanoPore platforms. Such work would be instrumental in improving detection of SL *trans*-splicing in any species.

Trichuris muris

T. muris is a gastrointestinal parasite of mice and is an important model system for studying mammalian gastrointestinal parasitism. It belongs to the same clade as *T. spiralis* and *P. punctatus* [70]. Comparative work with *T. spiralis* has identified a repertoire of 13 *T. muris* SLs, of which *Tmu*-SL1, *Tmu*-SL4 and *Tmu*-SL12 show structural similarity with *C. elegans* SL2 and are *trans*-spliced to the downstream gene of a bicistronic operon that is conserved among several nematode species [59]. However, the genome-wide landscape of SL *trans*-splicing and operons in *T. muris* is unresolved [59]. Since we thus cannot compare SLOPPR operon predictions against reference operons, we aimed to merely test the hypothesis that the three putative SL2-type SLs show genome-wide SL *trans*-splicing patterns that are distinct from those of the other ten SLs.

We downloaded genome assembly PRJEB126.WBPS15 from WormBase and five unstranded 2x100 bp libraries from bioproject PRJEB1054. We supplied SLOPPR with the 13 known SLs and designated *Tmu*-SL1, *Tmu*-SL4 and *Tmu*-SL12 as SL2-type. SLOPPR detected a relatively high SL *trans*-splicing rate of 35 % and clustered the 13 SLs into the expected groups comprising *Tmu*-SL1, *Tmu*-SL4 and *Tmu*-SL12

versus all other SLs (Figure 3). Using a relaxed SL2:SL1 threshold of two, SLOPPR predicted 718 operons comprising 1,492 operonic genes with a median intercistronic distance of 418 bp. This is larger than the c. 100 bp in *C. elegans* but is consistent with the observed elevated intercistronic distance among manually curated *Tmu* benchmark operons [59] and also with the considerably elevated intergenic distance among non-operonic genes (5,519 bp compared with c. 3,500 bp observed in all other species in this study; Supplemental Table S2). These results echo those we obtained in *T. spiralis* [19] and demonstrate that SLOPPR allows to identify subfunctionalisation among SLs that may correspond to SL1 and SL2-type *trans*-splicing. *Tmu*-SL1, *Tmu*-SL4 and *Tmu*-SL12 are very likely used to resolve polycistronic RNAs in this organism and SLOPPR has predicted plausible candidate operons that warrant future curation efforts.

Validation of SLOPPR in other eukaryotes

Having established that SLOPPR is a powerful method for predicting operons in organisms that use specialised SLs to resolve downstream operonic genes (SL2-type SLs), we finally aimed to illustrate that SLOPPR is also able to infer operons in organisms that lack SL specialisation. Here we demonstrate this ability in two tunicates, *Ciona intestinalis* and *Oikopleura dioica*, both of which have only a single SL that resolves operons but is also added to monocistronic genes.

In such situations, the SL must be designated as SL2-type such that all genes that receive the SL are classed as operonic; this set of genes will contain *bona fide* operonic genes but will also contain all monocistronic genes that receive the SL. Therefore, these initial candidate operonic genes must be filtered by intercistronic distance to partition out true operonic genes. SLOPPR can be configured to either use a user-supplied cutoff if the expected intercistronic distances are known, or to bisect the distribution of intercistronic distances empirically into two groups using K-means clustering and retaining those genes with short distances. By exploring several parameter combinations, specificity and sensitivity in partitioning out operons can be traded off (Figure 4).

Ciona intestinalis

The tunicate *C. intestinalis* splices a single SL to downstream operonic genes and infrequently to upstream operonic genes, consistent with operons in nematodes [21, 39]. Using short intergenic distances (<100 bp) as the sole criterion, a total of 1,310 operons comprising 2,909 genes have been predicted [21, 39]. These operons are predominantly dicistronic and have extremely small intercistronic distances, often lacking an intercistronic region altogether [21, 39], similar to the rare SL1-dependent operons observed in *C. elegans* [71]. The genome annotations take SL *trans*-splicing into account and define gene boundaries correctly between poly(A) and *trans*-splicing sites [21].

We obtained the KH genome assembly, the KH gene models (2013) and KH operon annotations (2013; containing 1,328 operons) from the Ghost database (http://ghost.zool.kyoto-u.ac.jp/download_kh.html). We used the same 13 RNA-Seq libraries from bioproject PRJNA376667 that contained disproportionately more evidence of SL *trans*-splicing using SLIDR than other tested libraries. SLOPPR detected an overall SL *trans*-splicing rate of 51 %, close to the 58 % expectation [65], although the libraries varied considerably in genome alignment rate and SL *trans*-splicing rate as observed earlier (Supplemental Table S2). We first ran SLOPPR with the same default configuration as for the nematodes, ignoring intercistronic distances after operon inference. Indeed, this first run predicted a vastly inflated set of 3,594 operons, of which 1,196 (33 %) matched reference operons and 90 % of the reference operons were detected. The contamination

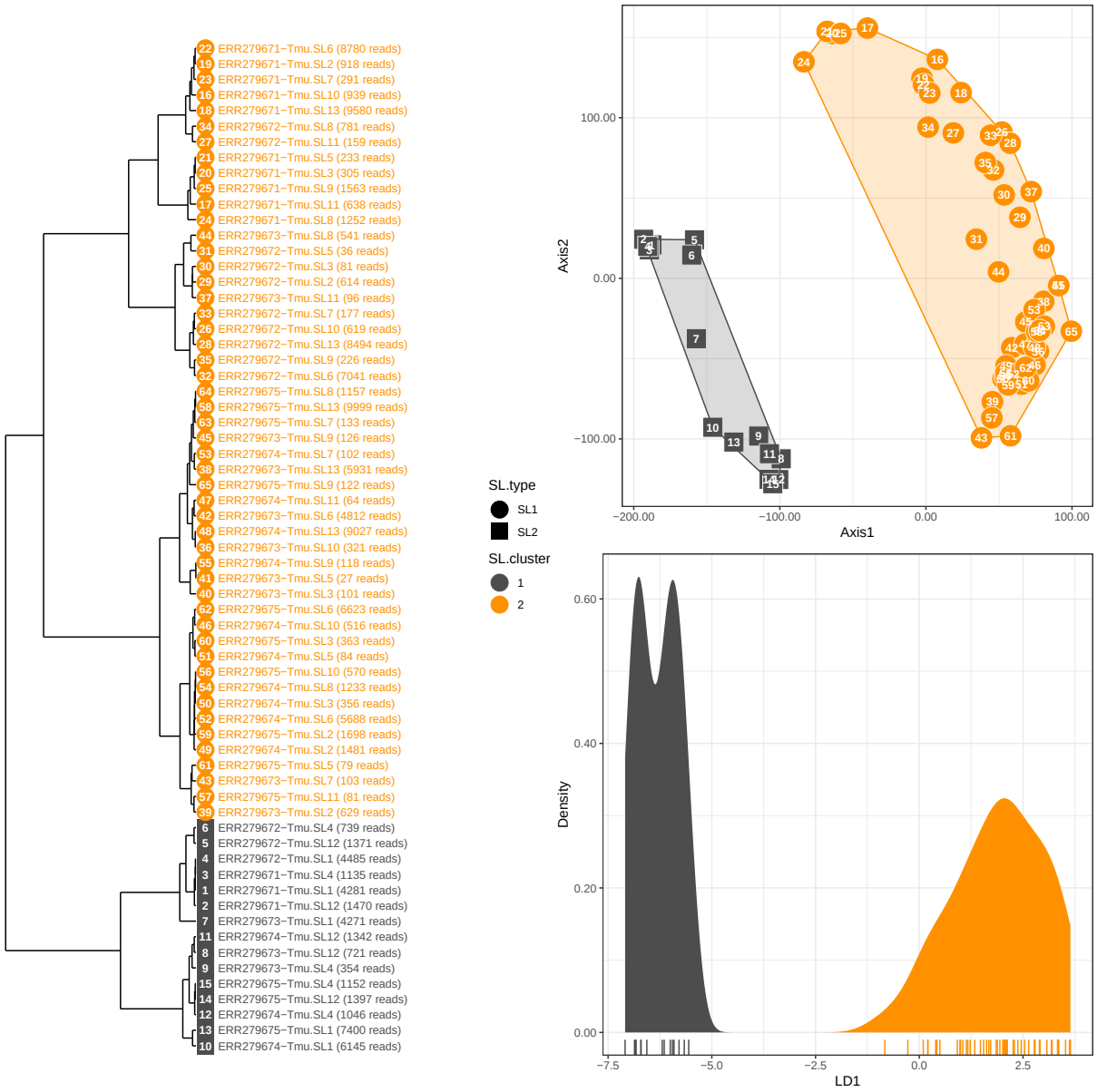

Figure 3: Genome-wide spliced-leader (SL) *trans*-splicing patterns among 13 SLs and five RNA-Seq libraries in *Trichuris muris*. Top right: generalized PCA of normalised genome-wide SL read counts. Symbol shape represents *a priori* SL type (circle: SL1; square: SL2) and colour represents cluster membership inferred via K-means clustering (dark grey: cluster1; orange: cluster2). Numbers inside symbols refer to library identifiers as detailed in the dendrogram on the left (hierarchical Ward's clustering of PCA eigenvectors). Bottom right: linear discriminant analysis between the two clusters, highlighting complete cluster differentiation. *Tmu*-SL1, *Tmu*-SL4 and *Tmu*-SL12 are correctly identified as distinct from all other SLs, confirming their functional specialisation as SL2-type SLs.

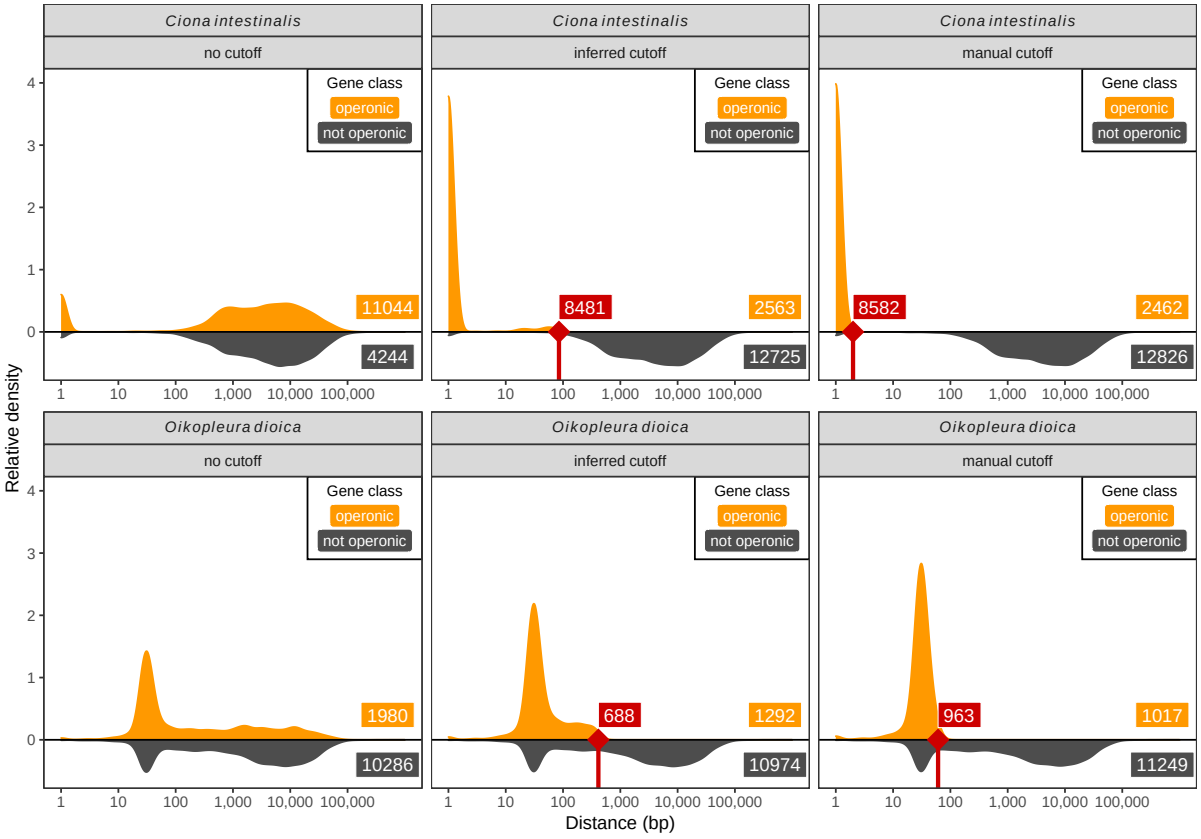

Figure 4: Separation of operonic genes from monocistronic genes in the absence of specialised spliced leaders (SLs), illustrated with SLOPPR data from the tunicates *Ciona intestinalis* (top panels) and *Oikopleura dioica* (bottom panels). All panels display distributions of distances between operonic and non-operonic genes, and labels provide gene numbers. Left panels: in both organisms, a single SL is added to monocistronic and operonic genes, causing SLOPPR to incorrectly designate monocistronic SL-receiving genes with large intergenic distances as operonic. Middle panels: an optimal distance cutoff for operonic genes is inferred via K-means clustering, and genes at or above the cutoff (red notches at 85 and 414 bp respectively) are re-classified as monocistronic non-operonic (red labels). Right panels: a lower manual cutoff (1 bp and 60 bp respectively; red notches at 2 and 61 bp) further reduces the set of genes retained as operonic. Note the peak of tightly-spaced non-operonic genes in the *O. dioica* panels; these genes are likely operonic genes but no SL evidence was obtained from the RNA-Seq data.

with monocistronic genes inflated the intercistronic distances (median 2,287 bp) but a distinct set of genes had very low intercistronic distances, likely representing true operons (Figure 4).

We then partitioned out true operonic genes by re-running SLOPPR with automatic inference of the optimal intercistronic distance cutoff and also switching off addition of upstream operonic genes. These strict settings require all genes in an operon to be SL *trans*-spliced and all operons to be at least dicistronic. SLOPPR predicted only 856 operons with a median intercistronic distance of 1 bp (inferred cutoff: 68 bp). Of these operons, 823 (96%) matched reference operons, indicating high specificity, but only 62 % of the reference operons were detected (Supplemental Table S2). Therefore, we re-ran the same analysis but re-instating upstream genes (i. e., allowing them not to be SL *trans*-spliced). This time, SLOPPR predicted 1,172 operons, of which 1,100 (94%) matched reference operons and represented a considerably improved 83 % of reference operons. The median intercistronic distance was again 1 bp (inferred cutoff: 84 bp), consistent with the notion that many operons in this organism have no intercistronic regions [21, 39]. To quantify this proportion of operons, we re-ran the analysis enforcing a maximum intercistronic distance of 1 bp. This yielded 1,128 operons, indicating that only 44 operons had intercistronic regions (Supplemental Table S2).

Overall, SLOPPR predicted most of the previously proposed operons and also novel operons comprising up to seven genes instead of six as previously reported [21]. This analysis demonstrates the flexibility that SLOPPR brings to operon prediction, by first identifying SL *trans*-splicing patterns and then filtering candidate operons by intercistronic distance. This approach enables extraction of operon sets with varying stringency and biological characteristics.

Oikopleura dioica

Like *C. intestinalis*, the tunicate *O. dioica* possesses only a single SL that is *trans*-spliced to both monocistronic genes and genes in operons, where upstream genes are not required to be SL *trans*-spliced [20, 72]. At least 39 % of genes are SL *trans*-spliced and 58 % of SL *trans*-spliced transcripts originate from operons [72]. A total of 1,765 operons comprising 5,005 genes have been predicted via short intercistronic distances of at most 60 bp [49]. We downloaded genome assembly GCA_000209535.1 (V3) and genome annotations from OikoBase (http://oikoarrays.biology.uiowa.edu/Oiko/), and operon annotations from the Genoscope Oikopleura Genome Browser (https://wwwdev.genoscope.cns.fr/oikopleura/). Curiously, both the genome and operon annotations also contain entries that reference an inferior, much smaller, assembly containing much shorter contigs (GCA_000209555.1). After removing these redundant entries from both annotation sets we were left with the expected 1,765 instead of 2,971 operons.

We downloaded four unstranded 2x90 bp libraries from bioproject PRJNA269316 and 16 stranded 2x100 bp libraries from bioproject PRJDB5668, representing various life stages. Similarly to *C. intestinalis*, we observed poor and highly variable background alignment rates (17-69 %) but large numbers of SL reads (5.8 million in total). However, these reads covered only 9 % of genes, which is much lower than the expected 39 % (Supplemental Table S2). In default mode, SLOPPR predicted 885 operons with median intercistronic distance of 57 bp. Of these, 644 (73 %) matched reference operons. As in *C. intestinalis*, the operons were contaminated with monocistronic genes having much larger intercistronic distances (median of 2,178 bp) (Figure 4). We therefore re-ran SLOPPR with inference of the optimal intercistronic distance cutoff and obtained 577 operons, of which 521 (90 %) matched reference operons. The median intercistronic distance was reduced to 33 bp, but the inferred cutoff was still fairly high at 413 bp (Figure 4). We thus re-ran the analysis with the same hard cutoff of 60 bp that was used to predict the 1,765 reference operons [49] and were left with 464 operons (median intercistronic distance of 31 bp), of which 454 (98%) matched reference operons (Supplemental Table S2). We also tested the

effect of enforcing SL *trans*-splicing at upstream genes (-u option) across the same three analysis runs and obtained a much more stringent set of 111-165 operons of which 106-143 (87-95 %) matched reference operons (Supplemental Table S2).

These results indicate that SLOPPR can discriminate operonic genes from monocistronic genes receiving the same SL, identify the vast majority of previously described operons and also predict a small number of novel operons. One limitation with this dataset was that only few genes received the SL, which was a consistent observation across all libraries tested from several bioprojects and suggests that more RNA-Seq data would need to be generated to fully characterise the SL *trans*-splicing landscape in this organism.

Conclusions

We have created two computational pipelines that fill a long-standing gap in our ability to identify and quantify SL *trans*-splicing and eukaryotic operons in any species where RNA-Seq data and reference genomes/transcriptomes are available. SLIDR is a more sensitive, specific and efficient SL discovery pipeline than SLFinder [33], able to uncover a wealth of untapped SL diversity. SLOPPR is the first universal pipeline to predict operons from SL *trans*-splicing events, closing this important gap left by existing SL quantification pipelines [34, 35]. We have demonstrated here and elsewhere [19] that SLOPPR identifies both *bona fide* and novel operons, blazing the trail for routine operon prediction in any organism with SL *trans*-splicing. Importantly, SLOPPR exploits biological replicates to infer subfunctionalisation among SLs and to moderate noise in SL quantification, which lays a foundational framework for developing a new field of eco-evolutionary "SL-omics" investigating differential SL usage and *trans*-splicing levels among biological replicates, experimental groups or wild populations.

A fundamental limitation of both SLIDR and SLOPPR is that they were designed for traditional RNA-Seq data where sequencing error is low but only a small fraction of reads originate from the 5' end of the transcript containing the SL. Most RNA-Seq library preparation methods also show considerable loss of coverage at the 5' end, which often limits SL detection to a short c. 10 bp portion at typically <1 % of reads [19, 35, 73]. This means that SLOPPR in particular is likely to underestimate the extent of SL *trans*-splicing and operonic gene organisation unless huge amounts of sequencing data are available [34] or specialised SL-enrichment library preparation methods are used [18, 22, 29]. However, our SLIDR analysis on *Hydra vulgaris* vividly demonstrates that SLs at nearly 100 % of all genes can be detected from RNA-Seq data if coverage is sufficient.

We decided to build these pipelines on RNA-Seq data because a wealth of datasets already exists for many species, which continues to grow rapidly. We are thus, for the first time, in the position to investigate SL *trans*-splicing systematically throughout the tree of life without needing to generate novel sequence data. Nevertheless, a powerful future avenue for capturing the full 5' end of transcripts is direct RNA or cDNA sequencing on the Oxford NanoPore or PacBio long-read platforms [74, 75]. This would require much less sequencing effort because the full molecule is sequenced instead of a short random fraction. SLIDR and SLOPPR could easily be expanded to accept long-read data but would require tailored error-tolerant screening methods to accommodate the higher error rate of NanoPore reads. As these long-read transcriptomics datasets become more commonplace, we expect SL-omics to become a routine molecular tool for uncovering the causes and consequences of this enigmatic source of molecular diversity.

Availability and requirements

Project name: SLIDR and SLOPPR

Project home page: https://github.com/wenzelm/slidr-sloppr

Operating system(s): Linux

Programming language: BASH, R

Other requirements: CUTADAPT (tested v2.3), GFFREAD (tested v0.11.4), HISAT2 (tested v2.1.0), BOWTIE2 (tested v2.3.5), SAMTOOLS (tested v1.9), BEDTOOLS (tested v2.28.0), SEQTK (tested v1.3), CDHIT (tested v4.8.1), VSEARCH (tested v2.4.3), BLASTN (tested v2.9.0), SUBREAD (tested v1.6.2), VIENNARNA (tested v2.4.14), R (tested v3.6.0), R packages *glmpca, data.table, parallel, ggdendro, MASS, ggplot2*

License: Creative Commons Attribution License (CC BY 4.0)

Any restrictions to use by non-academics: None

Abbreviations

SL spliced leader

UTR untranslated region

CPM counts per million reads

PCA principal component analysis

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All datasets analysed in this study are publicly available from NCBI (https://www.ncbi.nlm.nih.gov/), ENA (https://www.ebi.ac.uk/ena/browser/home), WormBase (https://wormbase.org/), Ghost database (http://ghost.zool.kyoto-u.ac.jp/download_kh.html), OikoBase (http://oikoarrays.biology.uiowa.edu/Oiko/) and Genoscope (https://wwwdev.genoscope.cns.fr/oikopleura/). Accession numbers are detailed in the main text and in supplemental materials. All data generated in this study are included in this published article and its supplemental material.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Biotechnology and Biological Sciences Research Council [BB/J007137/1 to JP and BM, and BB/T002859/1 to BM and JP]. The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

JP and BM acquired funding for the project, conceived the research and managed the research activity. MW designed and implemented the computational pipelines with major contributions by JP. MW carried out all data analyses and prepared all display items. MW wrote the manuscript with major contributions by JM and BM. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Bernadette Connolly for helpful discussions and Andreea Marin for testing the pipelines. The authors acknowledge the support of the Maxwell computer cluster funded by the University of Aberdeen.

References

[1] Nicholas A Stover, Michelle S Kaye, and Andre RO Cavalcanti. Spliced leader trans-splicing. *Current Biology*, 16(1):R8–R9, 2006.

[2] Erika L Lasda and Thomas Blumenthal. Trans-splicing. *Wiley Interdisciplinary Reviews: RNA*, 2 (3):417–434, 2011.

[3] Richard E Sutton and John C Boothroyd. Evidence for trans splicing in trypanosomes. *Cell*, 47(4): 527–535, 1986.

[4] Bart Cuypers, Malgorzata A Domagalska, Pieter Meysman, Géraldine de Muylder, Manu Vanaerschot, Hideo Imamura, Franck Dumetz, Thomas Wolf Verdonckt, Peter J Myler, Gowthaman Ramasamy, et al. Multiplexed spliced-leader sequencing: A high-throughput, selective method for rna-seq in trypanosomatids. *Scientific reports*, 7(1):1–11, 2017.

[5] George Cherian Pandarakalam, Michael Speake, Stuart McElroy, Ammar Alturkistani, Lucas Philippe, Jonathan Pettitt, Berndt Müller, and Bernadette Connolly. A high-throughput screen for the identification of compounds that inhibit nematode gene expression by targeting spliced leader trans-splicing. *International Journal for Parasitology: Drugs and Drug Resistance*, 10:28–37, 2019.

[6] Kenneth EM Hastings. Evolutionary origin of sl-addition trans-splicing: still an enigma. *Trends in genetics*, 21(4):240–247, 2005.

[7] Vassilis Douris, Maximilian J Telford, and Michalis Averof. Evidence for multiple independent origins of trans-splicing in metazoa. *Molecular biology and evolution*, 27(3):684–693, 2010.

[8] Romain Derelle, Tsuyoshi Momose, Michael Manuel, Corinne Da Silva, Patrick Wincker, and Evelyn Houliston. Convergent origins and rapid evolution of spliced leader trans-splicing in metazoa: insights from the ctenophora and hydrozoa. *RNA*, 16(4):696–707, 2010.

[9] Mainá Bitar, Mariana Boroni, Andréa Mara Macedo, Carlos Renato Machado, and Glória Regina Franco. The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. *Frontiers in genetics*, 4:199, 2013.

[10] Thomas Blumenthal. Operons in eukaryotes. *Briefings in Functional Genomics*, 3(3):199–211, 2004.

[11] John Spieth, Glenn Brooke, Scott Kuersten, Kristi Lea, and Thomas Blumenthal. Operons in c. elegans: polycistronic mrna precursors are processed by trans-splicing of sl2 to downstream coding regions. *Cell*, 73(3):521–532, 1993.

[12] Thomas Blumenthal. Trans-splicing and polycistronic transcription in caenohabditis elegans. *Trends in Genetics*, 11(4):132–136, 1995.

[13] Thomas Blumenthal, Donald Evans, Christopher D Link, Alessandro Guffanti, Daniel Lawson, Jean Thierry-Mieg, Danielle Thierry-Mieg, Wei Lu Chiu, Kyle Duke, Moni Kiraly, et al. A global analysis of caenorhabditis elegans operons. *Nature*, 417(6891):851, 2002.

[14] Kwang-Zin Lee and Ralf J Sommer. Operon structure and trans-splicing in the nematode pristionchus pacificus. *Molecular biology and evolution*, 20(12):2097–2103, 2003.

[15] David B Guiliano and Mark L Blaxter. Operon conservation and the evolution of trans-splicing in the phylum nematoda. *PLoS Genet*, 2(11):e198, 2006.

[16] Neale Harrison, Andreas Kalbfleisch, Bernadette Connolly, Jonathan Pettitt, and Berndt Müller. Sl2-like spliced leader rnas in the basal nematode prionchulus punctatus: New insight into the evolution of nematode sl2 rnas. *RNA*, 16(8):1500–1507, 2010.

[17] Bora Uyar, Jeffrey SC Chu, Ismael A Vergara, Shu Yi Chua, Martin R Jones, Tammy Wong, David L Baillie, and Nansheng Chen. Rna-seq analysis of the c. briggsae transcriptome. *Genome research*, 22(8):1567–1580, 2012.

[18] Amit Sinha, Claudia Langnick, Ralf J Sommer, and Christoph Dieterich. 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(9):1386–1397, 2014.

[19] Marius Wenzel, Christopher Johnston, Berndt Müller, Jonathan Pettitt, and Bernadette Connolly. Resolution of polycistronic rna by sl2 trans-splicing is a widely-conserved nematode trait. *RNA*, pages rna–076414, 2020.

[20] Philippe Ganot, Torben Kallesøe, Richard Reinhardt, Daniel Chourrout, and Eric M Thompson. Spliced-leader rna trans splicing in a chordate, oikopleura dioica, with a compact genome. *Molecular and cellular biology*, 24(17):7795–7805, 2004.

[21] Yutaka Satou, Katsuhiko Mineta, Michio Ogasawara, Yasunori Sasakura, Eiichi Shoguchi, Keisuke Ueno, Lixy Yamada, Jun Matsumoto, Jessica Wasserscheid, Ken Dewar, et al. Improved genome assembly and evidence-based global gene model set for the chordate ciona intestinalis: new insight into intron and operon populations. *Genome biology*, 9(10):R152, 2008.

[22] Mariana Boroni, Michael Sammeth, Sandra Grossi Gava, Natasha Andressa Nogueira Jorge, Andréa Mara Macedo, Carlos Renato Machado, Marina Moraes Mourão, and Glória Regina Franco. Landscape of the spliced leader trans-splicing mechanism in schistosoma mansoni. *Scientific Reports*, 8(1):1–14, 2018.

[23] Ferdinand Marlétaz, André Gilles, Xavier Caubit, Yvan Perez, Carole Dossat, Sylvie Samain, Gabor Gyapay, Patrick Wincker, and Yannick Le Parco. Chaetognath transcriptome reveals ancestral and unique features among bilaterians. *Genome biology*, 9(6):R94, 2008.

[24] Jarrod A Chapman, Ewen F Kirkness, Oleg Simakov, Steven E Hampson, Therese Mitros, Thomas Weinmaier, Thomas Rattei, Prakash G Balasubramanian, Jon Borman, Dana Busam, et al. The dynamic genome of hydra. *Nature*, 464(7288):592–596, 2010.

[25] Christian Preußer, Nicolas Jaé, and Albrecht Bindereif. mrna splicing in trypanosomes. *International Journal of Medical Microbiology*, 302(4-5):221–224, 2012.

[26] Huan Zhang, Yubo Hou, Lilibeth Miranda, David A Campbell, Nancy R Sturm, Terry Gaasterland, and Senjie Lin. Spliced leader rna trans-splicing in dinoflagellates. *Proceedings of the National Academy of Sciences*, 104(11):4618–4623, 2007.

[27] Huan Zhang and Senjie Lin. Retrieval of missing spliced leader in dinoflagellates. *PLoS One*, 4(1): e4129, 2009.

[28] Scott William Roy. Genomic and transcriptomic analysis reveals spliced leader trans-splicing in cryptomonads. *Genome biology and evolution*, 9(3):468–473, 2017.

[29] Mitsuhiro Matsuo, Atsushi Katahata, Soichirou Satoh, Motomichi Matsuzaki, Mami Nomura, Kenichiro Ishida, Yuji Inagaki, and Junichi Obokata. Characterization of spliced leader trans-splicing in a photosynthetic rhizarian amoeba, paulinella micropora, and its possible role in functional gene transfer. *PloS one*, 13(7):e0200961, 2018.

[30] Jonathan Pettitt, Berndt Müller, Ian Stansfield, and Bernadette Connolly. Spliced leader trans-splicing in the nematode trichinella spiralis uses highly polymorphic, noncanonical spliced leaders. *RNA*, 14(4):760–770, 2008.

[31] Yuelong Guo, David McK Bird, and Dahlia M Nielsen. Improved structural annotation of protein-coding genes in the meloidogyne hapla genome using rna-seq. In *Worm*, volume 3, page e29158. Taylor & Francis, 2014.

[32] Peter D Olson, Alan Tracey, Andrew Baillie, Katherine James, Stephen R Doyle, Sarah K Buddenborg, Faye H Rodgers, Nancy Holroyd, and Matt Berriman. Complete representation of a tapeworm genome reveals chromosomes capped by centromeres, necessitating a dual role in segregation and protection. *bioRxiv*, 2020.

[33] Javier Calvelo, Hernán Juan, Héctor Musto, Uriel Koziol, and Andrés Iriarte. Slfinder, a pipeline for the novel identification of splice-leader sequences: a good enough solution for a complex problem. *BMC bioinformatics*, 21(1):1–18, 2020.

[34] Nicolas J Tourasse, Jonathan RM Millet, and Denis Dupuy. Quantitative rna-seq meta analysis of alternative exon usage in c. elegans. *Genome research*, pages gr–224626, 2017.

[35] Carlo Yague-Sanz and Damien Hermand. Sl-quant: a fast and flexible pipeline to quantify spliced leader trans-splicing events from rna-seq data. *GigaScience*, 7(7):giy084, 2018.

[36] Timothy W Nilsen. Evolutionary origin of sl-addition trans-splicing: still an enigma. *Trends in Genetics*, 17(12):678 – 680, 2001. ISSN 0168-9525. doi: https://doi.org/10.1016/S0168-9525(01)02499-4. URL `http://www.sciencedirect.com/science/article/pii/S0168952501024994`.

[37] Daehwan Kim, Ben Langmead, and Steven L Salzberg. Hisat: a fast spliced aligner with low memory requirements. *Nature methods*, 12(4):357, 2015.

[38] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357, 2012.

[39] Yutaka Satou, Makoto Hamaguchi, Keisuke Takeuchi, Kenneth EM Hastings, and Nori Satoh. Genomic overview of mrna 5'-leader trans-splicing in the ascidian ciona intestinalis. *Nucleic acids research*, 34(11):3378–3388, 2006.

[40] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[41] Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. Vsearch: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016.

[42] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10(1): 421, 2009.

[43] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[44] Jeffrey D Thomas, Richard C Conrad, and Thomas Blumenthal. The c. elegans trans-spliced leader rna is bound to sm and has a trimethylguanosine cap. *Cell*, 54(4):533–539, 1988.

[45] Stacey N Barnes, Rick E Masonbrink, Thomas R Maier, Arun Seetharam, Anoop S Sindhu, Andrew J Severin, and Thomas J Baum. Heterodera glycines utilizes promiscuous spliced leaders and demonstrates a unique preference for a species-specific spliced leader over c. elegans sl1. *Scientific reports*, 9(1356), 2019.

[46] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6(1):26, 2011.

[47] Donald Evans and Thomas Blumenthal. trans splicing of polycistroniccaenorhabditis elegans pre-mrnas: Analysis of the sl2 rna. *Molecular and cellular biology*, 20(18):6659–6667, 2000.

[48] R Core team. R: A language and environment for statistical computing. 2013.

[49] France Denoeud, Simon Henriet, Sutada Mungpakdee, Jean-Marc Aury, Corinne Da Silva, Henner Brinkmann, Jana Mikhaleva, Lisbeth Charlotte Olsen, Claire Jubin, Cristian Cañestro, et al. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science*, 330 (6009):1381–1385, 2010.

[50] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

[51] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2013.

[52] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511, 2010.

[53] Mary Ann Allen, LaDeana W Hillier, Robert H Waterston, and Thomas Blumenthal. A global analysis of c. elegans trans-splicing. *Genome research*, 21(2):255–264, 2011.

[54] F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology*, 20(1): 1–16, 2019.

[55] F Krueger. Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files*, 2015.

[56] Donald W Nelson and Barry M Honda. Two highly conserved transcribed regions in the 5s dna repeats of the nematodes caenorhabditis elegans and caenorhabditis briggsae. *Nucleic acids research*, 17(21):8657–8667, 1989.

[57] Lincoln D Stein, Zhirong Bao, Darin Blasiar, Thomas Blumenthal, Michael R Brent, Nansheng Chen, Asif Chinwalla, Laura Clarke, Chris Clee, Avril Coghlan, et al. The genome sequence of caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biol*, 1(2):e45, 2003.

[58] Nadine Borchert, Christoph Dieterich, Karsten Krug, Wolfgang Schütz, Stephan Jung, Alfred Nordheim, Ralf J Sommer, and Boris Macek. Proteogenomics of pristionchus pacificus reveals distinct proteome structure of nematode models. *Genome research*, 20(6):837–846, 2010.

[59] Jonathan Pettitt, Lucas Philippe, Debjani Sarkar, Christopher Johnston, Henrike Johanna Gothe, Diane Massie, Bernadette Connolly, and Berndt Müller. Operons are a conserved feature of nematode genomes. *Genetics*, 197(4):1201–1211, 2014.

[60] Debjani Sarkar. *Spliced leader trans-splicing and operons in Dorylaimida (Nematoda)*. PhD thesis, University of Aberdeen, 2014.

[61] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644, 2011.

[62] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

[63] Peng Cui, Qiang Lin, Feng Ding, Chengqi Xin, Wei Gong, Lingfang Zhang, Jianing Geng, Bing Zhang, Xiaomin Yu, Jin Yang, et al. A comparison between ribo-minus rna-sequencing and polya-selected rna-sequencing. *Genomics*, 96(5):259–265, 2010.

[64] Anna R Dahlgren, Erica Y Scott, Tamer Mansour, Erin N Hales, Pablo J Ross, Theodore S Kalbfleisch, James N MacLeod, Jessica L Petersen, Rebecca R Bellone, and Carrie J Finno. Comparison of poly-a+ selection and rrna depletion in detection of lncrna in two equine tissues using rna-seq. *Non-coding RNA*, 6(3):32, 2020.

[65] Jun Matsumoto, Ken Dewar, Jessica Wasserscheid, Graham B Wiley, Simone L Macmil, Bruce A Roe, Robert W Zeller, Yutaka Satou, and Kenneth EM Hastings. High-throughput sequence analysis of ciona intestinalis sl trans-spliced mrnas: alternative expression modes and gene function correlates. *Genome research*, 20(5):636–645, 2010.

[66] Amanda E Vandenberghe, Thomas H Meedel, and Kenneth EM Hastings. mrna 5'-leader trans-splicing in the chordates. *Genes & development*, 15(3):294–303, 2001.

[67] Brendan Yeats, Jun Matsumoto, Sandra I Mortimer, Eiichi Shoguchi, Nori Satoh, and Kenneth EM Hastings. Sl rna genes of the ascidian tunicates ciona intestinalis and ciona savignyi. *Zoological science*, 27(2):171–180, 2010.

[68] Nicholas A Stover and Robert E Steele. Trans-spliced leader addition to mrnas in a cnidarian. *Proceedings of the National Academy of Sciences*, 98(10):5693–5698, 2001.

[69] Aleksandar Rajkovic, Richard E Davis, J Neil Simonsen, and FRITZ M RoTTMAN. A spliced leader is present on a subset of mrnas from the human parasite schistosoma mansoni. *Proceedings of the National Academy of Sciences*, 87(22):8879–8883, 1990.

[70] Ashleigh B Smythe, Oleksandr Holovachov, and Kevin M Kocot. Improved phylogenomic sampling of free-living nematodes enhances resolution of higher-level nematode phylogeny. *BMC evolutionary biology*, 19(1):121, 2019.

[71] Carol Williams, Lei Xu, and Thomas Blumenthal. Sl1 trans splicing and 3'-end formation in a novel class of caenorhabditis elegans operon. *Molecular and Cellular Biology*, 19(1):376–383, 1999. ISSN 0270-7306. doi: 10.1128/MCB.19.1.376. URL `https://mcb.asm.org/content/19/1/376`.

[72] Gemma B Danks, Martina Raasholm, Coen Campsteijn, Abby M Long, J Robert Manak, Boris Lenhard, and Eric M Thompson. Trans-splicing and operons in metazoans: translational control in maternally regulated development and recovery from growth arrest. *Molecular biology and evolution*, 32(3):585–599, 2015.

[73] Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, 38(12):e131–e131, 2010.

[74] Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, et al. Highly parallel direct rna sequencing on an array of nanopores. *Nature methods*, 15(3):201, 2018.

[75] Daniel P Depledge, Kalanghad Puthankalam Srinivas, Tomohiko Sadaoka, Devin Bready, Yasuko Mori, Dimitris G Placantonakis, Ian Mohr, and Angus C Wilson. Direct rna sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nature communications*, 10(1): 754, 2019.