OXFORD

Briefings in Bioinformatics, 2023, **24(1)**, 1–9

https://doi.org/10.1093/bib/bbac610

Problem Solving Protocol

slORFfinder: a tool to detect open reading frames resulting from *trans*-splicing of spliced leader sequences

Bo Song[†], Hao Li[†], Mengyun Jiang, Zhongtian Gao, Suikang Wang, Lei Gao, Yunsheng Chen and Wujiao Li

Corresponding authors. Yunsheng Chen, Department of Laboratory Medicine, Shenzhen Children's Hospital, Shenzhen 518038, China, E-mail: chenyunshenglw@163.com; Wujiao Li, Department of Laboratory Medicine, Shenzhen Childrens' Hospital, Shenzhen 518038, China, E-mail: hnnd059@gmail.com.

[†]Bo Song and Hao Li contributed equally.

Abstract

Trans-splicing of a spliced leader (SL) to the 5′ ends of mRNAs is used to produce mature mRNAs in several phyla of great importance to human health and the marine ecosystem. One of the consequences of the addition of SL sequences is the change or disruption of the open reading frames (ORFs) in the recipient transcripts. Given that most SL sequences have one or more of the trinucleotide NUG, including AUG in flatworms, *trans*-splicing of SL sequences can potentially supply a start codon to create new ORFs, which we refer to as slORFs, in the recipient mRNAs. Due to the lack of a tool to precisely detect them, slORFs were usually neglected in previous studies. In this work, we present the tool slORFfinder, which automatically links the SL sequences to the recipient mRNAs at the *trans*-splicing sites identified from SL-containing reads of RNA-Seq and predicts slORFs according to the distribution of ribosome-protected footprints (RPFs) on the *trans*-spliced transcripts. By applying this tool to the analyses of nematodes, ascidians and euglena, whose RPFs are publicly available, we find wide existence of slORFs in these taxa. Furthermore, we find that slORFs are generally translated at higher levels than the annotated ORFs in the genomes, suggesting they might have important functions. Overall, this study provides a tool, slORFfinder (https://github.com/songbo446/slORFfinder), to identify slORFs, which can enhance our understanding of ORFs in taxa with SL machinery.

Keywords: spliced leader, trans-splicing, ribosome footprints, open reading frames

Introduction

Trans-splicing of spliced leader (SL) sequences has been found in seven taxa that are widely distributed phylogenetically, including nematodes, flatworms, cnidarians, rotifers, ascidians, dinoflagellates and euglenozoa [1, 2]. Many organisms in these taxa, such as trypanosomes, flatworms and nematodes, are pathogenic to humans. Others, such as dinoflagellates, are crucial to the marine ecosystem as endosymbionts of coral or causal algae of harmful blooms.

The process of SL *trans*-splicing is similar to the process of intron splicing. The SL RNA molecule is separated into two parts by the splice donor site, possibly a GT dinucleotide. The 5′ part is the leader sequence, which will be added to the recipient mRNAs, and the 3′ part contains a putative binding site of Sm-protein [1]. A two-step process was proposed for the *trans*-splicing of SLs [1, 3]. The 3′ part is spliced and forms a 'Y' shape structure with the outron sequences resembling the lariat structure required by the splicing of the intron, which is then followed by the ligation of the 5′ SL sequence to the acceptor sites on the transcripts [1, 3]. A recent study suggests that the trimming of the 5′ SL might also be involved in the maturation of the SL *trans*-spliced mRNAs [4]. Eventually, *trans*-splicing of SL results in the replacement of a part of the sequences of the original 5′ untranslated regions (5′ UTRs) on the transcripts, which can have a variety of biological consequences. For example, it can stabilize the mRNAs [5], remove regulatory elements, such as upstream open reading frames (uORFs) in the 5′ UTR, and enhance the translation efficiency of recipient mRNAs [6]. It can also result in the translation of alternative ORFs or disruption of existing ORFs on the transcripts [7]. The *trans*-splicing of the flatworm SL sequence, which ends with the trinucleotide AUG at its 3′ terminus, can even create new ORFs by providing a start codon to the recipient transcripts [8] (Figure 1A). Given that the SL sequences in some species also carry at least one cognate start codon (NUG), providing start codons to the recipient transcripts could be pervasive in these species.

Bo Song is an associate professor of Chinese Academy of Agricultural Sciences. His research interests include the development of bioinformatic pipelines, comparison and analysis of omics' datasets.
Hao Li is a student of Shenzhen University. He is interested in bioinformatics.
Mengyun Jiang is a graduate student of Chinese Academy of Agricultural Sciences and Henan University, China. She is interested in the development of bioinformatic pipelines.
Zhongtian Gao is a student of Shenzhen University. She is interested in bioinformatics.
Suikang Wang is an associate professor of Chinese Academy of Agricultural Sciences. His research interest mainly focuses on the epigenetic dynamics and proteomics during development.
Lei Gao is an associate professor of Shenzhen University. His research focuses on the analyses and mining of omics' datasets.
Yunsheng Chen is a professor in Shenzhen Children's Hospital. He is interested in laboratory diagnosis of hematological diseases.
Wujiao Li is a senior technician in Shenzhen Children's Hospital. His research focuses on genomic comparison and development of bioinformatic pipelines.

Received: August 31, 2022. **Revised:** November 16, 2022. **Accepted:** December 11, 2022

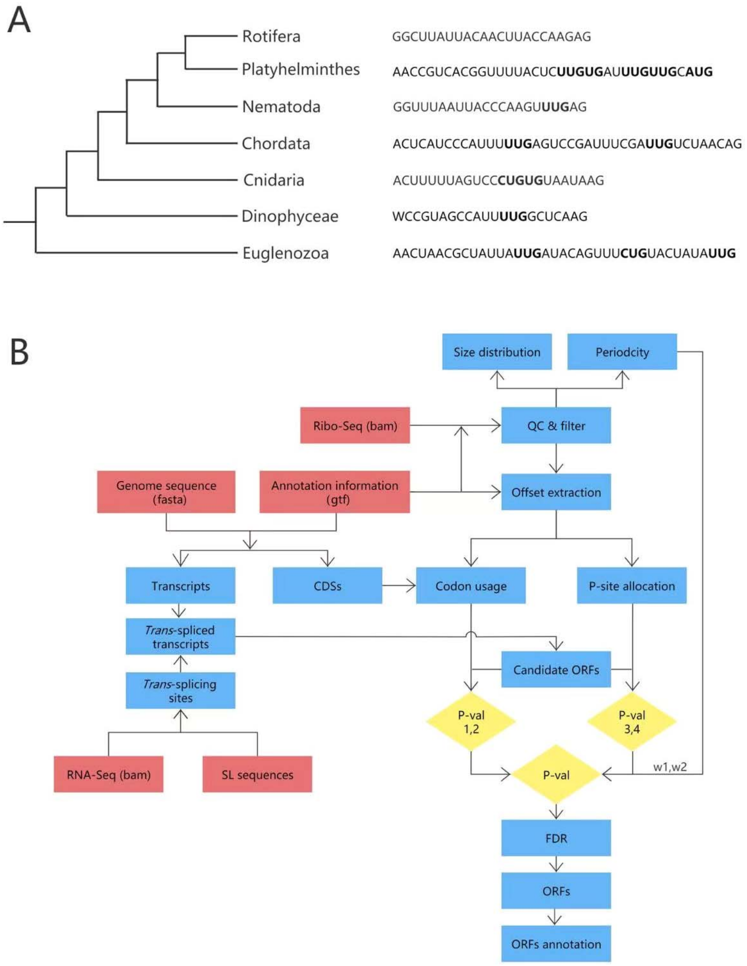

Figure 1. Overview of slORFfinder. (**A**) SL sequences in different taxa. The bolder letters represent the cognate start codons. (**B**) The workflow of slORFfinder. Four input files and SL sequences, indicated in red in the workflow, are used for slORFfinder: the reference genome sequence (in fasta format), the genome annotation (in gtf format), the alignment of RNA-Seq reads and Ribo-Seq reads (in bam format), and the sequence of SL. The genome sequences and annotations are used to extract the sequences of nascent mRNAs, the alignments of RNA-Seq reads are used to identify the SL *trans*-splicing sites according to the SL sequences provided, and the alignments of Ribo-Seq reads are used to identify the P-sites on mature mRNAs to predict ORFs.

Following the development and application of the Ribo-Seq technique, accumulating evidence is showing the roles of small/short ORFs, particularly uORFs, in various organisms. Ribo-Seq, also called ribosome profiling, captures ribosome-protected mRNA fragments (RPFs), so it can be used to predict ORFs [9]. Many tools implementing different algorithms have been developed to predict ORFs from RPFs, such as riboHMM [10], RiboTaper [11], RiboCode [12], RiboWave [13], ORFquant [14] and RiboNT [15]. However, most, if not all, of the existing tools were developed to study mammals or yeasts, which do not possess SL *trans*-splicing machinery. All existing tools for detecting ORFs along the native transcripts will certainly miss slORFs. Due to the lack of appropriate tools, the investigation of slORFs has been neglected in previous studies.

In this work, we developed slORFfinder, a tool that automatically detects the splicing sites of SL sequences and predicts ORFs on the *trans*-spliced mature transcripts. By applying slORFfinder to the datasets of *Caenorhabditis elegans*, *Columnea brenneri*, *C. remanei* [16], *Typanosoma brucei* [17] and *Oikopleura dioica* [18], we found that slORFs are pervasive in these species. Furthermore, the slORFs are generally more actively translated and have higher translation efficiencies in comparison to other types of ORFs, including the annotated ORFs (aORFs).

Methods and datasets

Datasets

To detect slORFs, we searched for reads of RPFs in studies of species with SL machinery by querying the keywords 'RiboSeq,' 'RPF,' 'Translatome,' and 'Ribosome profiling,' coupled with the scientific name of the corresponding species or higher-level taxon. As a result, we found RPFs for euglenozoa *Trypanosoma brucei* [17] and *Trigonoscuta cruzi* [19]; nematodes *C. elegans*, *C. brenneri* and *C. remanei* [16]; a marine chordate *O. dioica* [18]; and a dinoflagellate *Lingulodinium polyedra* [20]. The RPFs and parallelized RNA-Seq reads of these species were downloaded from the NCBI database, except *L. polyedra* due to the lack of reference genome sequences and annotations (Table S1). We evaluated the quality of the RPFs and discarded those of poor quality, without significant 3-nt periodicity which is required for the identification of ORFs. As a result, only the RPFs of *C. elegans*, *C. brenneri*, *C. remanei*, *O. dioica* and *T. brucei* were retained for further analyses.

Alignment of reads

The adaptors were trimmed from the Ribo-Seq reads, resulting in RPFs, which were then mapped to the corresponding reference genomes (Table S2) using STAR [21] with default parameters. The

RNA-Seq reads were also aligned to the references to enable the detection of the SL splicing sites.

Simulation of testing datasets

To measure the accuracy and sensitivity of slORFfinder, we produced datasets containing simulated slORFs with different expression levels. To eliminate the impact from real SL *trans*-splicing events, the simulated datasets were generated from a previously published RNA-Seq and Ribo-Seq dataset of rice [22], in which the SL *trans*-splicing machinery is absent, and an imagined SL sequence (CCGUAGCCAUUUUGGCUCAG) was used to create the simulated slORFs. We selected genes with expression levels ranging from 1 to more than 800 reads per kilobase per million mapped reads (RPKMs) and replaced their reads with the simulated reads from their corresponding slORFs, which were created by replacing the first 3, 6 or 9 nucleotides with the SL sequence. The simulated reads were produced using ART [23] with the reads length set to 50 bp according to the reads features in the original datasets. Then slORFfinder was used on the simulated dataset to recover the simulated slORFs, and the recall rate (number of recovered slORFs/total number of simulated ORFs) was calculated. This test was repeated three times independently.

Quantification of translation levels

The translation and transcription levels of ORFs were calculated using featureCounts [24]. For the slORFs overlapped with the annotated major ORFs (mORFs), the reads mapped to the overlapped regions were proportionally assigned to the slORFs or mORFs according to the ratio of reads mapped to the distinct regions in slORFs or annotated mORFs before they were collectively used to calculate the translation or transcription levels of slORFs or mORFs.

Validation of predicted slORFs using MS datasets

To validate the predicted slORFs, we searched for protein mass spectrometry (MS) datasets of the tested species. Only the MS data of *C. elegans* were found in the PRIDE archive (https://www.ebi.ac.uk/pride/). We downloaded the MS dataset of *C. elegans* (PXD032260) to validate the slORFs identified in this dataset. The raw data of MS were loaded into MaxQuant [25] to search for evidence of slORF-encoded peptides with default parameters.

Results

Overview of slORFfinder

Our proposed tool, slORFfinder, predicts ORFs primarily based on the mapping positions of RPFs, and it also considers the codon usages like some other tools [15, 26]. It consists of three steps and takes four input files, namely the reference genome sequence (in fasta format), the genome annotation (in gtf format), the alignment of RNA-Seq reads and Ribo-Seq reads (in bam format), and the sequence of SL (Figure 1B). Like many other tools, slORFfinder first calculates the offsets of RPFs of different lengths and the genome-wide usages of codons according to their appearance in the annotated gene coding sequences. Second, it reads the alignment of RNA-Seq reads to identify SL splicing sites from the soft clipped reads by the following two criteria: (1) more than eight bases (by default) are soft clipped from the aligned reads, and (2) these clipped sequences are the 3′ terminuses of the inputted SL sequence(s). The sequences upstream of the SL splicing sites are then replaced by the SL sequence to produce the sequences of mature transcripts. Third, slORFfinder reads the mapping loci of RPFs and allocates them to the positions on the mature transcripts according to the offsets trained in the first step to predict ORFs following the algorithm implemented in RiboNT [15] or OrfPP [26]. Briefly, this tool automatically tests if the depths of RPFs and the usages of codons at the n^{th} positions are significantly greater than at the $(n+1)^{\mathrm{th}}$ and $(n+2)^{\mathrm{th}}$ positions, where n denotes the positions that are multiples of three in the candidate ORFs, and it combines the P values in these four tests into a final P value, taking into account the weights assigned according to the quality of RPFs (Figure 1B) [15]. We denote *trans*-ORFs resulting from the *trans*-splicing of SL as slORFs to distinguish them from the *trans*-ORFs resulting from the fusion of exons from different genes. It is optional to predict the ORFs in all the transcripts or only the slORFs in the *trans*-spliced transcripts.

Detection of non-canonical start codons

Our tool, slORFfinder, is designed based upon the assumption that the cognate start codons, such as UUG, on SL sequences can initiate translation of the *trans*-spliced transcripts, but there is no direct evidence to support this assumption. To confirm this assumption, we searched for the footprints at the initiation sites in the RPFs. The SL sequences are added to the 5′ terminus of the recipient mRNAs with varied distances from the start codons. Usually, the SL sequences are not translated when they are distant from the translation initiation sites, whereas those close to the start codons can be captured in the RPFs. Therefore, the RPFs containing SL sequences (SL-RPFs) represent the footprints of the translation initiation complex, allowing the identification of the utilized start codons. We calculated the offsets from the start codons to the 5′ terminuses of the RPFs of different sizes by performing metagene analyses. Briefly, the RPFs mapped to the annotated start codons were pooled and plotted (Figure 2A), which enabled the calculation of the distances from the 5′ terminuses of the RPFs to the start codons, because the most upstream RPFs must be derived from the ribosomes stalled at the initiation sites. The data showed that the distances from the 5′ terminuses of the RPFs to the translating codons were consistently 12 nt for *C. elegans*, *C. brenneri* and *C. remanei* (Figure 2A). By searching the triplets at 12 nt of the SL-RPFs in the datasets of these species, we found that many non-canonical start codons other than AUG, such as UUG, were also used to initiate translation (Figure 2B). As expected, AUG was the most frequently used start codon, followed by UUG, in all three species, with usages ranging from 19% to 22% (Figure 2C). Together, these results suggest the capability of translation initiation of cognate start codons.

Test of slORFfinder in simulated datasets

To evaluate the performance of slORFfinder, we artificially created a dataset including a total of 1753 simulated slORFs of varying translation levels ranging from 1 to more than 800 RPKM. By applying slORFfinder to this simulated dataset, we recovered 145–191 of these slORFs with a recall rate of 8.27–9.87%. The results reveal that the identification of slORFs is remarkably affected by their abundance. Clearly, the higher the translation level of slORFs, the greater the chance they can be recovered by slORFfinder (Figure 3). As much as 81.81% of the slORFs at translation levels higher than 800 RPKM were successfully recovered, whereas only 7.45% of the slORFs at lower levels (1–200 RPKM) were recovered (Figure 3). We noticed that the recall rate increased dramatically at the level of 200–400 RPKM, suggesting slORFfinder might be powerless to detect slORFs with translation levels lower than 200 RPKM. These results suggest slORFfinder can efficiently identify slORFs expressed at adequate levels.

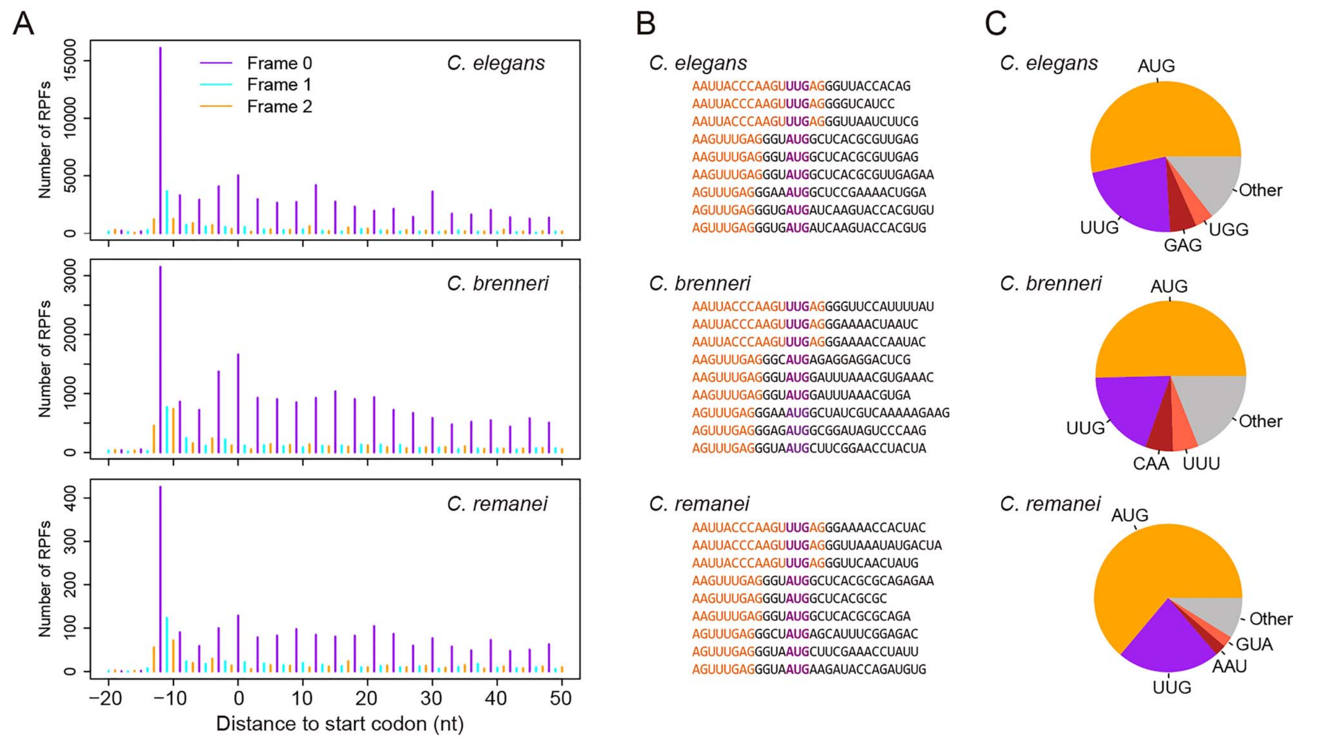

Figure 2. Evidence of the use of non-canonical start codons. (**A**) The offsets from the 5′ terminuses of RPFs to the translating codons calculated from metagene plots of *C. elegans*, *C. brenneri* and *C. remanei*. (**B**). Examples of footprints at translation initiation sites showing the use of non-canonical start codons. (**C**) The usage of codons at translation initiation sites.

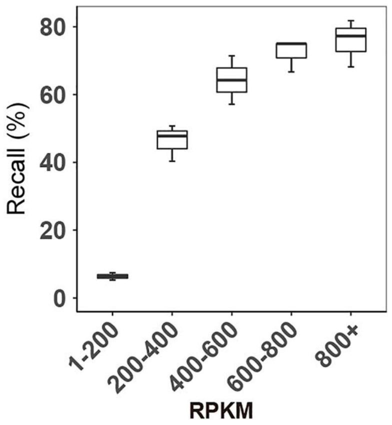

Figure 3. The recall rate of slORFfinder tested in simulated datasets.

The identification of slORFs

There are seven taxa of species that are known to have SL *trans*-splicing machinery [2] (Figure 1A). We searched for their RPFs in NCBI datasets (Table S1) to predict slORFs in these species. As a result, RPFs of seven species were found, including three nematodes, two euglenozoas, one chordate and one dinoflagellate. The RPFs of the dinoflagellate *L. polydera* were not used for slORF detection because of the lack of reference genome sequence. To better identify the slORFs in their genomes, we pooled the RNA-Seq reads and Ribo-Seq reads separately, which were then mapped to the corresponding reference genomes using STAR [21]. The alignment files were then used to predict slORFs by slORFfinder with corresponding SL sequences inputted. Thus, we identified 334, 276, 233 and 586 slORFs from *C. elegans*, *C. brenneri*, *O. dioica* and *T. brucei*, respectively (Table S3), suggesting the common presence of slORFs in these species. Several of them are translated in frames different from their corresponding host ORFs. Specifically, the translating frames of 8.38% (28 of 334), 8.34% (23 of 276), 4.72% (11 of 233) and 1.88% (11 of 586) of the slORFs are shifted (Table S4). The RPFs of several slORFs are shown in Figure 4, in which the cyan, orange and purple lines indicate the RPFs at frames 0, 1 and 2 in the transcript, respectively, where frame 0 corresponds to slORFs. These plots confirm that the majority of the RPFs are in the same frame in these predicted ORFs. Several slORFs are overlapped with the downstream annotated mORFs. Some of the overlapped slORFs are translated in a frame different from that of the annotated mORFs (Figure 4). Figure 4 shows some slORFs translated out of or in the frame of their corresponding annotated mORFs. These plots clearly show that both the slORFs and mORFs are actively translated under the tested conditions, and the counts of RPFs from slORFs are generally much higher than the counts of RPFs from mORFs, suggesting higher translation levels of slORFs in these studies. Furthermore, some slORFs are in the same frame as the mORFs. They usually share the same stop codons with the mORFs but with varied initiation sites due to the *trans*-splicing of SL sequences (Table S3).

To validate these predicted slORFs, we searched for protein MS data in public databases, but found only the MS of *C. elegans*. We queried the MS data against the slORF-encoded peptides and found that 29.64% of the slORFs were supported by the MS data (Table S3). Some slORFs may have gone undetected due to their low levels of expression. Indeed, the MS-supported slORFs were expressed at higher levels than those not supported (Figure S1). Although only approximately one-third of the predicted slORFs were supported by MS evidence, this ratio is adequate given the limited power of MS in detecting proteins encoded by inactive genes. Indeed, only 6.98% of the annotated ORFs in the

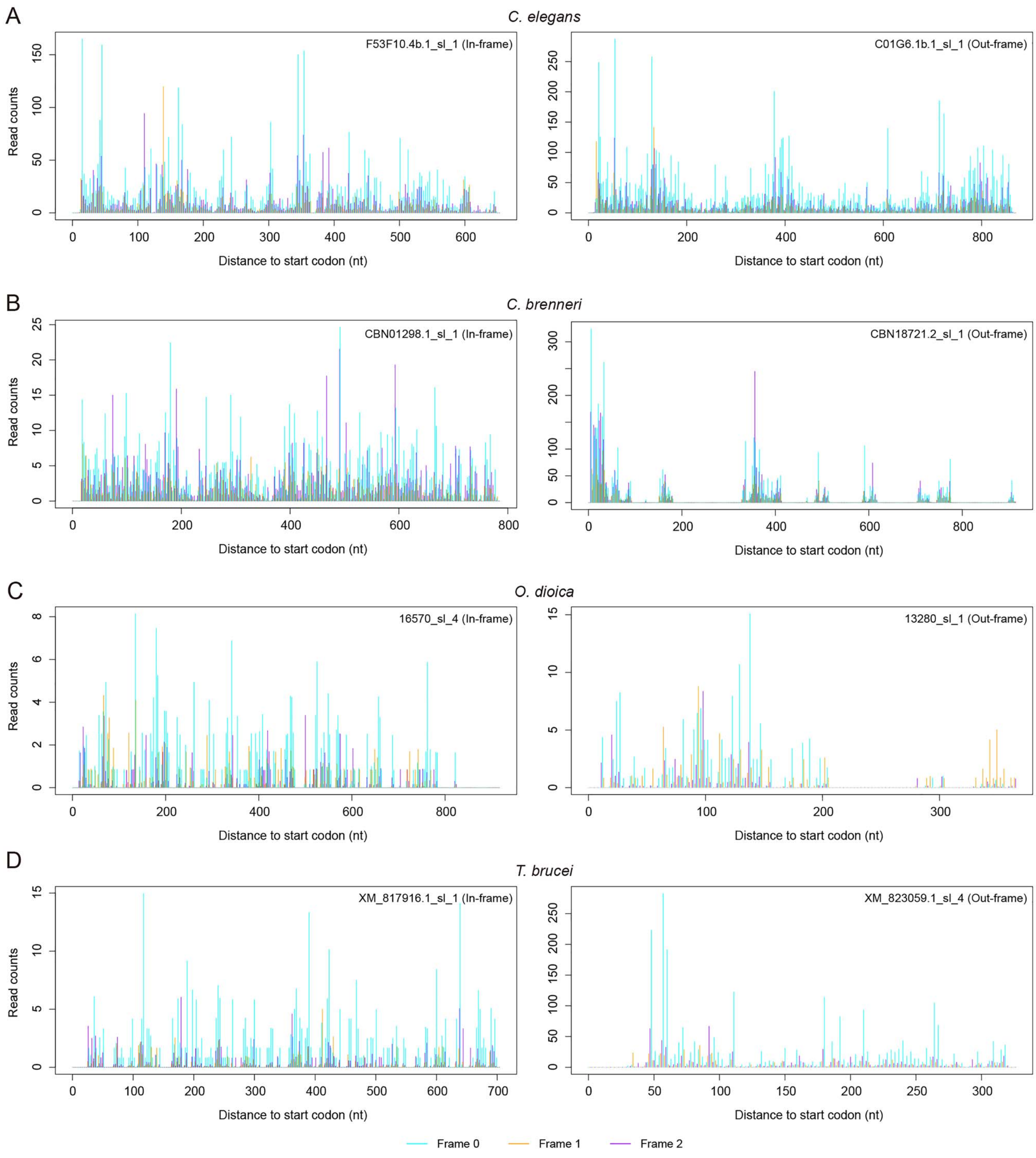

Figure 4. The distribution of RPFs in some examples of the slORFs identified in (**A**) *C. elegans*, (**B**) *C. brenneri*, (**C**) *O. dioica* and (**D**) *T. brucei*. The cyan, orange and purple lines represent frames 0, 1 and 2, respectively. The labels of 'out-frame' and 'in-frame' in the parentheses in each plot indicate whether these slORFs are translated in a frame different from its corresponding annotated ORF (out-frame) or not (in-frame).

C. elegans genome were supported by this MS dataset. Among the *C. elegans* slORFs with shifted frames, 28.57% (8 of 28) were supported by MS data. This percentage is close to the overall supporting degree (29.64%) of the *C. elegans* slORFs, suggesting that the confidence of these frame-shifted slORFs is comparable to those not shifted. These results suggest that the *trans*-splicing of SL can substantially increase the diversity of proteins in the cells.

The translation of slORFs

To compare the translation levels of slORFs and the levels of other types of ORFs, we collected the RNA-Seq reads and RPFs from the studies of *C. elegans*, *C. brenneri*, *T. brucei* and *O. dioica*, in which more than 32 different treatments, such as the *C. elegans* and *T. brucei* at different development stages (Table S1), were investigated. We first compared the translation levels and translation efficiencies between the mRNAs with or without slORFs (Figure 5). Our data

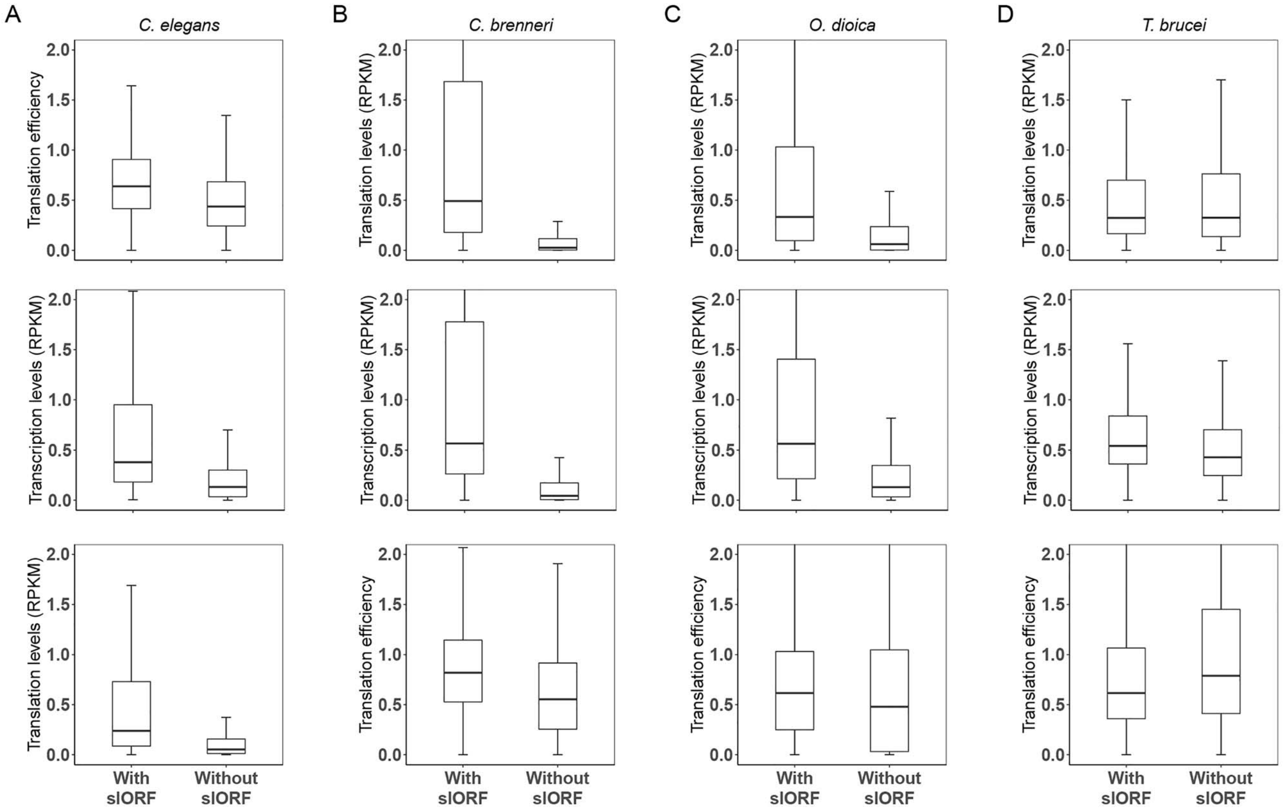

Figure 5. Comparison of the translation levels between the mRNAs with or without slORFs in (**A**) *C. elegans*, (**B**) *C. brenneri* (**C**) *O. dioica* and (**D**) *T. brucei*.

indicate that the transcripts with slORFs are more actively translated than those without slORFs in most of the tested species except *T. brucei* (Figure 5), in which both the translation levels and translation efficiencies are lower for the transcripts with slORFs. We also compared the translation levels and efficiencies of slORFs and the annotated ORFs by calculating their transcription and translation levels under different conditions. Our data indicate that the translation levels and translation efficiencies of slORFs are generally higher than those of other types of ORFs, including the annotated ORFs (Figure 6). The high levels of translation imply critical biological functions of slORFs.

Discussion

The *trans*-splicing of SL sequences is known to be influential to the translation of the recipient mRNAs. For example, the nematode SL contributes an N-2,2,7-trimethylguanosine cap (TMG-cap) to more than 70% of the recipient mRNAs, and both the TMG-cap and the SL itself are required for efficient translation of mRNAs [27]. SL addition can result in the truncation of ORFs, the alteration of signal peptides and the interruption of existing ORFs [7]. Researchers first noticed that the SL sequence might be able to provide a start codon to the recipient mRNAs in flatworms because its SL sequence ends with a trinucleotide AUG (Figure 1A). It was reported in 2006 that the 3′ terminus AUG in the SL sequence is in-frame with 28% of the *trans*-spliced transcripts of *Schistosoma mansoni*, and more than 40 cDNAs require the AUG from the SL sequence to initiate their translation in a correct frame [8]. The profiling of RPFs at the translation initiation sites in many organisms revealed the translation initiation by many non-canonical start codons, including UUG, CUG, GUG and other variants of AUG, such as AGG and ACG [28–30]. Given that the SL end with AUG can supply a start codon to the recipient mRNAs [8], and most SL sequences have cognate start codons (Figure 1A), it is reasonable to assume that the addition of SL sequences can potentially create new ORFs in the recipient mRNAs. Indeed, our analyses of SL-RPFs supported the fact that non-canonical start codons can also initiate translation in the tested species. By reanalyzing the previously published RPFs of nematodes, ascidians and euglena, we successfully identified slORFs in *C. elegans*, *C. brenneri*, *T. brucei* and *O. dioica*. The failure of slORF identification in *T. cruzi* and *C. remaneri* could be attributed to the small amount of RPFs. Together with the slORFs reported in the flatworm *S. mansoni*, our data suggest that slORFs are common in the taxa with SL machinery.

It is noteworthy the number of slORFs identified in this study is highly underestimated because this prediction depends on the total number of SL-containing reads, which are informative in identifying the *trans*-splicing sites. In this work, we used RNA-Seq reads to identify *trans*-splicing sites, but these reads encompassed only a small proportion of SL-containing reads. Furthermore, the limited data and poor quality of RPFs in some of the datasets also limited the identification of slORFs. Instead of normal RNA-Seq reads, large-scale identification of slORFs can use the SL-PCR method [31] to capture more SL *trans*-splicing sites before sequencing the libraries of cDNA. Briefly, reverse-transcribing the mRNAs with a customized primer designed according to the SL sequences and a random primer can, in principle, capture all the *trans*-splicing sites. This approach can efficiently enrich the SL *trans*-splicing sites in the datasets.

Trans-splicing is known to be involved in the translational regulation of genes. For example, it can repress the translation of nutrient-responsive genes under the challenged conditions [18, 32]. A study on *T. brucei* reported that more than 85% of the transcripts of the annotated protein-coding genes were

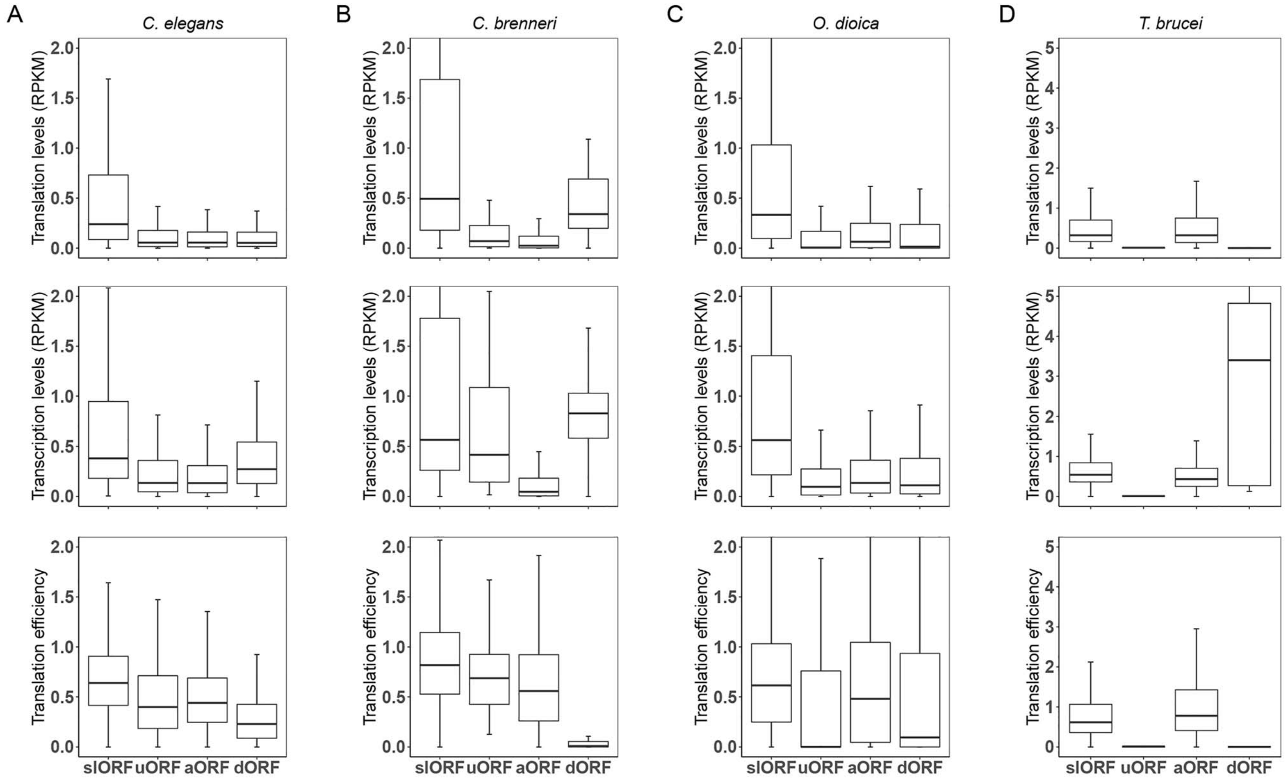

Figure 6. Comparison of the translation levels between slORFs and other types of ORFs in (**A**) *C. elegans*, (**B**) *C. brenneri*, (**C**) *O. dioica* and (**D**) *T. brucei*. uORF: upstream ORF, aORF: annotated ORF, dORF: downstream ORF.

trans-spliced, resulting in various changes of the existing ORFs, and 40% of the *trans*-spliced transcripts were dynamically regulated under different conditions [7]. Many *trans*-splicing events can affect the translation of the ORFs in the recipient mRNAs in a manner dependent on the *trans*-splicing sites. For example, events resulting in the trimming of AUG from coding sequences can lead to the failure of translation, while those SLs inserted into uORF can change the translation levels of the downstream ORFs, given that uORFs usually regulate the translation of downstream ORFs [6, 9, 33]. Due to the lack of a proper tool to identify slORFs in the transcriptome, the roles of slORFs are rarely investigated. Many of the slORFs identified in this study were overlapped with the annotated ORFs but in different frames, whereas some of them shared the same frame with the annotated ORFs but with varied lengths. *Trans*-splicing of SL to create new ORFs may play a significant part in increasing the diversity of the protein reservoir without any genomic changes. The relatively high translation levels of slORFs imply they have critical biological roles, but more research on this topic is necessary.

Overall, we present the tool slORFfinder to identify slORFs from Ribo-Seq reads in taxa with SL *trans*-splicing machinery. As slORFs are widely present in these taxa with SL machinery, as shown by previous studies and our data, this tool will help in our future studies of these species and substantially further our understanding of the roles of slORFs and SL *trans*-splicing.

Usage of slORFfinder

slORFfinder has been deposited in GitHub (https://github.com/songbo446/slORFfinder) and can be easily downloaded and installed in Linux. It requires four inputs: the reference genome sequence (in fasta format), the genome annotation (in GTF format), RNA-Seq alignments (in bam/sam format) and Ribo-Seq alignments (in bam/sam format), and a sequence or sequences of SL(s).

In some cases, more than one SL sequence needs to be inputted since some species have more than one SL sequence [2], and some species, such as dinoflagellate species [34], have degenerate bases in the SL sequence. slORFfinder allows the input of multiple SL sequences by separating these SLs using a comma without spaces. slORFfinder searches for the SL-containing reads in the inputted alignments of RNA-Seq reads and Ribo-Seq reads. Regardless of the length of the SL sequences, it searches for the last 8 bp by default at the 3′ end of the inputted SL(s) in the reads. Users can also customize this length by giving a value to the option '—slseed.' If the degenerate bases are located at the 5′ end of the SL sequence, or if multiple SL sequences share an identical sequence at their 3′ ends, users can input only the common sequence as long as all the potential start codons (NUG) are included.

The filtering, adaptor trimming, and mapping of RNA-Seq reads and Ribo-Seq reads should be performed before the use of this tool. slORFfinder reads the alignment files of RNA-Seq and Ribo-Seq in bam or sam format directly. STAR is the recommended aligner, but the outputs (in bam/sam format) of other aligners are also acceptable. The aligner should allow partial alignment of reads because slORFfinder looks for the soft clipped reads. If the partial aligned reads are not reported, the splicing sites cannot be determined, which will lead to the failure of slORF identification. It can be expected that longer reads will be more routinely used in transcriptomic studies to resolve alternative splicing and *trans*-splicing events. In these cases, users can also report the alignment results into 'sam' format before they are used to identify slORFs. For example, if minimap2 [35] is used for the alignment of long-reads, a parameter '-a' can be selected to output the results in 'sam' format.

Key Points

- slORFfinder, a tool to predict ORFs resulting from SL trans-splicing (slORFs), is developed.
- Evidence shows the wide presence of slORFs in the taxa with SL machinery.
- slORFfinder could facilitate large-scale identification of slORFs.

Author contributions

B. S., W. L and Y. C. conceived the work; B. S, H. L. and W. L. coded and tested the program; H. L., Z. G. and L. G. prepared and collected the data and performed the analyses; S. W. and H. L. prepared the simulated datasets; M. J. and S. W. participated in the verification of predicted ORFs. W. L. and M. J. participated in the visualization of data analyses. B. S., W. L., Y. C. and H. L. wrote the manuscript. Y. C. and W. L. revised the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (No. 31601042 to Bo Song) and the Agricultural Science and Technology Innovation Program, Guangdong High-level Hospital Construction Fund and supported by Guangdong Provincial High-level Hospital Construction Special Fund (ynkt2021-zz39), the Research Grants for Postdocs Settle in Shenzhen from Shenzhen Government to Wujiao Li and Bo Song.

References

1. Stover NA, Kaye MS, Cavalcanti ARO. Spliced leader trans-splicing. *Curr Biol* 2006;**16**:R8–9.
2. Bitar M, Boroni M, Macedo A, *et al.* The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. *Front Genet* 2013;**4**:199.
3. Michaeli S. Trans-splicing in trypanosomes: machinery and its impact on the parasite transcriptome. *Future Microbiol* 2011;**6**: 459–74.
4. Song Y, Zaheri B, Liu M, *et al.* Fugacium spliced leader genes identified from stranded RNA-Seq datasets. *Microorganisms* 2019;**7**:171.
5. Hastings KEM. SL trans-splicing: easy come or easy go? *Trends Genet* 2005;**21**:240–7.
6. Yang Y-F, Zhang X, Ma X, *et al.* Trans-splicing enhances translational efficiency in C. elegans. *Genome Res* 2017;**27**:1525–35.
7. Nilsson D, Gunasekera K, Mani J, *et al.* Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of *Trypanosoma brucei*. *PLoS Pathog* 2010;**6**:e1001037.
8. Cheng G, Cohen L, Ndegwa D, *et al.* The flatworm spliced leader 3′-terminal AUG as a translation initiator methionine∗. *J Biol Chem* 2006;**281**:733–43.
9. Ingolia NT, Ghaemmaghami S, Newman JRS, *et al.* Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 2009;**324**:218–23.
10. Raj A, Wang SH, Shim H, *et al.* Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* 2016;**5**:e13328.
11. Calviello L, Mukherjee N, Wyler E, *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods* 2016;**13**:165–70.
12. Xiao Z, Huang R, Xing X, *et al.* De novo annotation and characterization of the translatome with ribosome profiling data. *Nucleic Acids Res* 2018;**46**:e61–1.
13. Xu Z, Hu L, Shi B, *et al.* Ribosome elongating footprints denoised by wavelet transform comprehensively characterize dynamic cellular translation events. *Nucleic Acids Res* 2018;**46**:e109–9.
14. Calviello L, Hirsekorn A, Ohler U. Quantification of translation uncovers the functions of the alternative transcriptome. *Nat Struct Mol Biol* 2020;**27**:717–25.
15. Song B, Jiang M, Gao L. RiboNT: a noise-tolerant predictor of open reading frames from ribosome-protected footprints. *Life* 2021;**11**:701.
16. Stadler M, Fire A. Conserved translatome remodeling in nematode species executing a shared developmental transition. *PLoS Genet* 2013;**9**:e1003739.
17. Jensen BC, Ramasamy G, Vasconcelos EJR, *et al.* Extensive stage-regulation of translation revealed by ribosome profiling of Trypanosoma brucei. *BMC Genomics* 2014;**15**:911.
18. Danks GB, Galbiati H, Raasholm M, *et al.* Trans-splicing of mRNAs links gene transcription to translational control regulated by mTOR. *BMC Genomics* 2019;**20**:908.
19. Smircich P, Eastman G, Bispo S, *et al.* Ribosome profiling reveals translation control as a key mechanism generating differential gene expression in Trypanosoma cruzi. *BMC Genomics* 2015;**16**:443.
20. Bowazolo C, Song B, Dorion S, *et al.* Orchestrated translation specializes dinoflagellate metabolism three times per day. *Proc Natl Acad Sci USA* 2022;**119**:e2122335119.
21. Dobin A, Davis CA, Schlesinger F, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21.
22. Yang X, Song B, Cui J, *et al.* Comparative ribosome profiling reveals distinct translational landscapes of salt-sensitive and -tolerant rice. *BMC Genomics* 2021;**22**:612.
23. Huang W, Li L, Myers JR, *et al.* ART: a next-generation sequencing read simulator. *Bioinformatics* 2012;**28**:593–4.
24. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30.
25. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 2008;**26**: 1367–72.
26. Jiang M, Ning W, Wu S, *et al.* Three-nucleotide periodicity of nucleotide diversity in a population enables the identification of open reading frames. *Brief Bioinform* 2022;**23**:bbac210.
27. Lall S, Friedman CC, Jankowska-Anyszka M, *et al.* Contribution of trans-splicing, 5′ -leader length, cap-poly(a) synergism, and initiation factors to nematode translation in an Ascaris suum embryo cell-free system. *J Biol Chem* 2004;**279**: 45573–85.
28. Lee S, Liu B, Lee S, *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci USA* 2012;**109**:E2424–32.
29. Eisenberg AR, Higdon AL, Hollerer I, *et al.* Translation initiation site profiling reveals widespread synthesis of non-AUG-initiated protein isoforms in yeast. *Cell Syst* 2020;**11**: 145–160.e145.

30. Gelsinger DR, Dallon E, Reddy R, *et al.* Ribosome profiling in archaea reveals leaderless translation, novel translational initiation sites, and ribosome pausing at single codon resolution. *Nucleic Acids Res* 2020;**48**:5201–16.
31. Matsumoto J, Dewar K, Wasserscheid J, *et al.* High-throughput sequence analysis of Ciona intestinalis SL trans-spliced mRNAs: alternative expression modes and gene function correlates. *Genome Res* 2010;**20**:636–45.
32. Danks GB, Raasholm M, Campsteijn C, *et al.* Trans-splicing and operons in metazoans: translational control in maternally regulated development and recovery from growth arrest. *Mol Biol Evol* 2015;**32**:585–99.
33. Radío S, Garat B, Sotelo-Silveira J, *et al.* Upstream ORFs influence translation efficiency in the parasite *Trypanosoma cruzi*. *Front Genet* 2020;**11**:166.
34. Zhang H, Hou Y, Miranda L, *et al.* Spliced leader RNA trans-splicing in dinoflagellates. *Proc Natl Acad Sci USA* 2007;**104**: 4618–23.
35. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100.