Characterization of spliced leader trans-splicing in a photosynthetic rhizarian amoeba, Paulinella micropora, and its possible role in functional gene transfer
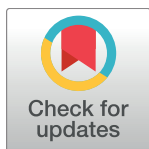
Characterization of spliced leader *trans*-splicing in a photosynthetic rhizarian amoeba, *Paulinella micropora*, and its possible role in functional gene transfer

Mitsuhiro Matsuo[1], Atsushi Katahata[1], Soichirou Satoh[1], Motomichi Matsuzaki[2], Mami Nomura[3], Ken-ichiro Ishida[3], Yuji Inagaki[4], Junichi Obokata[1]*

1 Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, Kyoto, Japan,
2 Department of Biomedical Chemistry, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan, **3** Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan, **4** Center for Computational Sciences, University of Tsukuba, Tsukuba, Japan

* Obokata@kpu.ac.jp

Abstract

Paulinella micropora is a rhizarian thecate amoeba, belonging to a photosynthetic *Paulinella* species group that has a unique organelle termed chromatophore, whose cyanobacterial origin is distinct from that of plant and algal chloroplasts. Because acquisition of the chromatophore was quite a recent event compared with that of the chloroplast ancestor, the *Paulinella* species are thought to be model organisms for studying the early process of primary endosymbiosis. To obtain insight into how endosymbiotically transferred genes acquire expression competence in the host nucleus, here we analyzed the 5′ end sequences of the mRNAs of *P. micropora* MYN1 strain with the aid of a cap-trapper cDNA library. As a result, we found that mRNAs of 27 genes, including endosymbiotically transferred genes, possessed the common 5′ end sequence of 28–33 bases that were posttranscriptionally added by spliced leader (SL) *trans*-splicing. We also found two subtypes of SL RNA genes encoded by the *P. micropora* MYN1 genome. Differing from the other SL *trans*-splicing organisms that usually possess poly(A)-less SL RNAs, this amoeba has polyadenylated SL RNAs. In this study, we characterize the SL *trans*-splicing of this unique organism and discuss the putative merits of SL *trans*-splicing in functional gene transfer and genome evolution.

Introduction

Splice leader (SL) *trans*-splicing is an RNA maturation process that adds a short leader sequence of 16–52 bases to the 5′ end of precursor mRNAs [1]. This process was first reported for *Trypanosoma* [2, 3] and later found in diverse organisms [1, 4]. SL *trans*-splicing occurs similarly to *cis*-splicing [5] and is mediated by the *trans*-spliceosome that contains U2, U4, U5, U6 snRNPs and SL snRNP, which is in place of U1 snRNP that is used for *cis*-spicing [6–11]. SL snRNP carries SL RNA to the 5′ end of the recipient precursor mRNAs [5, 12].

Physiological roles of the SL *trans*-splicing were intensively studied in trypanosomes and nematodes, revealing that it is essential to their viability [13, 14]. These organisms possess polycistronic gene clusters whose transcripts are converted to monocistronic mature mRNAs by SL *trans*-splicing [15–18]. SL *trans*-splicing can replace their unfavorable 5′ UTR sequences for translation with preferable ones [19–22].

One of the peculiarities of SL *trans*-splicing is its sporadic appearance in eukaryotic lineages. SL *trans*-splicing was reported in metazoan [23], Euglenozoa [2, 3, 24], dinoflagellates [25], Perkinsea [26], and Rhizaria [27] but not found in fungi, plants and most other protists [1]. In metazoan, SL *trans*-splicing is also patchily distributed; for example, tunicates in chordates [28, 29] and copepod and amphipod [23, 30] in arthropoda perform SL *trans*-splicing, but no evidence has been found in closely related vertebrates and insects [1]. The fragmented distribution should reflect multiple gains or losses of the SL *trans*-splicing system in the eukaryotic lineage [31], and repeated acquisitions of SL *trans*-splicing are suggested in metazoan evolution [23, 32]. Interestingly, from the observation that an U1 snRNA fragment can be converted into SL RNA by addition of only splice donor sequence and a few mutations [33], it is proposed that SL *trans*-splicing might have emerged by mutation of U snRNAs [1, 23]. SL *trans*-splicing might have easily emerged in eukaryote evolution.

Photosynthetic *Paulinella* species are rhizarian thecate amoebas with a unique photosynthetic organelle, namely, chromatophora, derived from the cyanobacterial ancestor distinct from that of plant and algal chloroplasts [34–36]. Its endosymbiosis is thought to have occurred about 100 million years ago [37]; hence, its chromatophore is a very young organelle compared with circa 1.9 billion-year-old chloroplasts [38]. Therefore, *Paulinella* species attracted the attention of researchers interested in the early process of primary endosymbiosis and endosymbiotic gene transfer (EGT) events [39, 40]. For these purposes, genome and transcriptome analyses [27] have been performed for the CCAC0185 strain of *Paulinella chromatophora* [36].

We have been interested in the molecular mechanisms of EGT and HGT (horizontal gene transfer) especially in the process of how transferred alien genes obtain expression competence in their host nucleus. In this respect, the promoter and 5′ UTR of the recently transferred genes may contain rich information. In this study, we analyzed the 5′ UTRs of mRNAs in the *P. micropora* MYN1 strain, a member of the photosynthetic *Paulinella* group [41] but phylogenetically distant from *P. chromatophora* [42]. We found that the nuclear genes of *P. micropora* MYN1, including those derived from EGT, are also at least in part subjected to SL *trans*-splicing, as implicated in *P. chromatophora* [27]. This study shows detailed characterization of SL *trans*-splicing of *P. micropora*, and we also discuss the potential merits of SL *trans*-splicing from the viewpoint of functional gene transfer and genome evolution.

Results

Detection of SL *trans*-splicing in *P. micropora* MYN1

To obtain information about the 5′ end sequences of *P. micropora* MYN1's transcripts, we sequenced the cap-trapper cDNA libraries [43] of this organism using the Ion PGM sequencer (Thermo Fisher Scientific, Waltham, MA, USA). After trimming the linker- and adapter-sequences, the reads of organelle transcripts and rRNAs were discarded. Finally, we obtained 325,863 reads (44 Mb) with the average read length of 136 bases. Mapping of these reads on the actin and polyubiquitin mRNAs ensured that the 5′ end sequences were enriched (S1 Fig). Because axenic culture has never been established for *P. micropora*, the sequence reads should include those from the contaminating bacteria. Therefore, we removed the reads that were highly similar to bacterial genomes. Furthermore, we collected the reads that encode N-

terminal 50 amino acid residues of 606 independent eukaryotic proteins in the public database (NCBI-nr) by BLASTX (e-value < 1e-5) to obtain a more reliable dataset, and subjected them to the following analyses.

The bioinformatics analysis of the obtained reads revealed that 16 distinct genes, including a *psbN* gene [40], contain a common 28–33 base sequence at their 5′ ends (Fig 1). This common sequence was confirmed by RT-PCR and Sanger sequencing of 8 genes, indicating that they are not sequencing artifacts.

It is known that the existence of 15–52 base common sequences at the 5′ ends of various mRNAs is a characteristic of SL *trans*-splicing. Furthermore, the 5′ end common sequences of *P. micropora* contain UUU triplet and had G nucleotide at the 3′ terminus (Fig 1), which are typical features of SL sequences shared by SL *trans*-splicing organisms known thus far [4].

To confirm that SL *trans*-splicing occurs in *P. micropora*, we analyzed an SL *trans*-spliced gene that encodes calmodulin (Fig 2). We cloned a calmodulin gene from −588 to +758 relative to ATG initiation codon and found that it contained two introns. SL sequence was not found in the genomic sequence, and instead a 3′ acceptor splicing motif, i.e., AG (Fig 2A and 2B, asterisk) was found at the predicted position. Structural organization of the genomic and cDNA sequences was compared using the PCR primers shown in Fig 2A (arrows a, b, c, and d). A primer pair a–d amplified a fragment only from cDNA but not from genomic DNA, indicating that the SL sequence was added to the mRNA by a posttranscriptional event (Fig 2C). To discriminate whether the SL sequence should be added by *trans*-splicing or long-range *cis*-splicing, we first determined the transcription start sites (TSSs) of this gene by cap-trapper cDNA sequencing, which revealed their locations at −151, −150, and −136 relative to the ATG codon (Fig 2B, thick arrows). On this basis, we designed primers b and c, which located downstream and upstream of the TSSs, respectively (Fig 2A). A primer pair b–d (Fig 2A) produced PCR fragments of the same size from both genomic and cDNA (Fig 2C), indicating that in our RNA preparations, primer b sequences (primer b consists of two nested primers underlined in Fig 2B) were retained only by the primary transcripts before splicing out the introns. Next, the primer pair c–d gave the amplified fragment only from the genomic DNA but not from cDNA, indicating that this gene does not produce long transcripts starting upstream of the above TSSs. These results reject the possibility that the SL sequence was added to the mRNA by long-range *cis*-splicing, and hence we conclude that SL *trans*-splicing occurs in this organism.
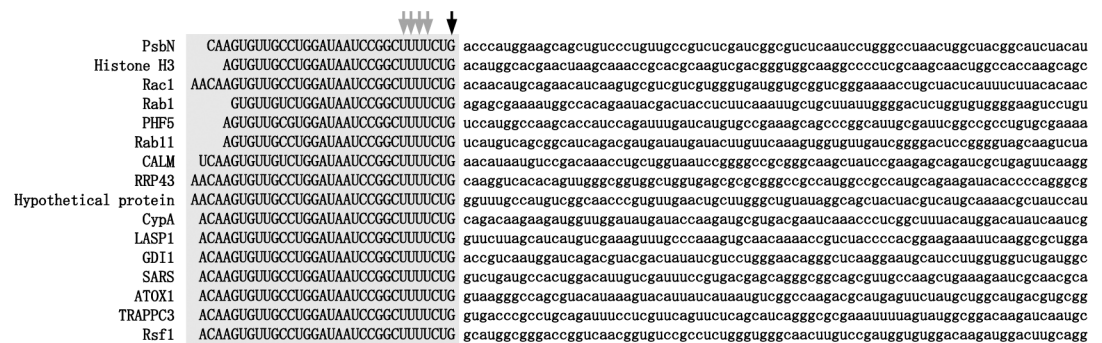
Fig 1. Alignment of the spliced leader *trans*-spliced mRNAs of *P. micropora*. cDNAs contain 28–33 base common sequences at their 5′ ends. These contigs were annotated using COG and NCBI-nr database (PsbN; photosystem II PsbN, Histone H3, Rac1; Ras-related protein Rac1, Rab1; GTP-binding protein, PHF5; PHF5-domain-containing protein, Rab11; GTPase Rab11/YPT3, CALM; Calmodulin, RRP43; Exosome complex exonuclease RRP43, CypA; Cyclophilin, LASP1; LIM and SH3 domain protein, GDI1; Rab GDP dissociation inhibitor alpha, SARS; seryl-tRNA synthetase, ATOX1; copper transport protein ATOX1, TRAPPC3; trafficking protein particle complex subunit 3, Rsf1; RNA-binding protein Rsf1). The common sequence is shaded, with the UUU triplet and 3′ terminal G indicated by the gray and black arrows, respectively.

https://doi.org/10.1371/journal.pone.0200961.g001
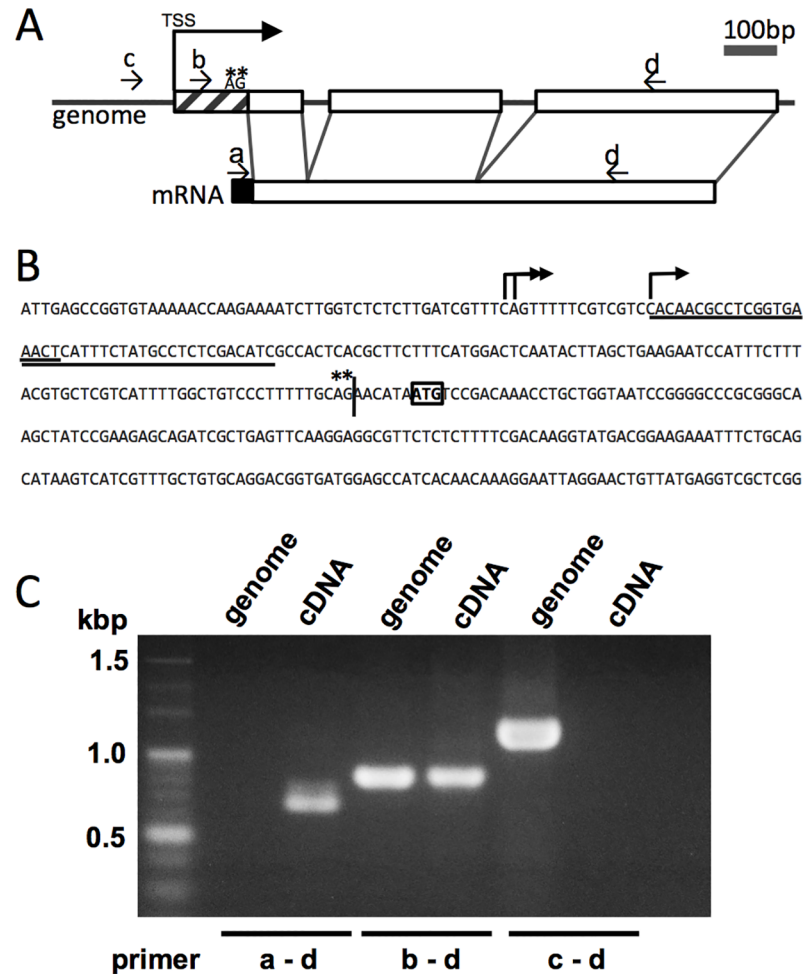
Fig 2. Structure of the calmodulin gene of *P. micropora*. A. Schematic model of the calmodulin gene and its mRNA. Primers used for PCR- and RT-PCR analyses are shown by small arrows a, b, c, and d. The outron and spliced leader are shown as hashed and black boxes, respectively. The thick arrow indicates the transcription start site (TSS). **B.** Genomic sequence of the calmodulin gene, with the ATG initiation codon boxed. The junction of the outron and exon is indicated by the vertical bar. TSS and the pyrimidine rich region are indicated by the thick arrow and dots, respectively. The underlines show the position of primer b (consisting of two nested primers). **C.** PCR profiles of the calmodulin gene using genomic and cDNA templates.

SL *trans*-splicing occurs for the genes of various metabolism in *P. micropora*

When we searched for the SL sequence shown in Fig 1, the cDNA sequence with reads similar to bacterial genes were discarded for reasons mentioned above. However, *P. micropora* undoubtedly has its own genes derived from cyanobacterial endosymbiont [39] and probably also from horizontally transferred genes from unidentified bacteria [40]. To salvage those nuclear genes, we collected whole cap-trapper cDNA reads that contain 20-base SL conserved sequence (TGGATAATCCGGCTTTTCTG), and we obtained 1204 reads, which were assembled into 773 contigs. Although these 5′ terminal sequences were generally too short to cover the protein-coding regions, BLASTX analysis detected 11 new protein genes. In conjunction with the genes listed in Fig 1, we detected 27 SL *trans*-spliced genes in total. Their functional classification showed that SL *trans*-splicing in *P. micropora* occurs for genes of diverse metabolisms

A

B
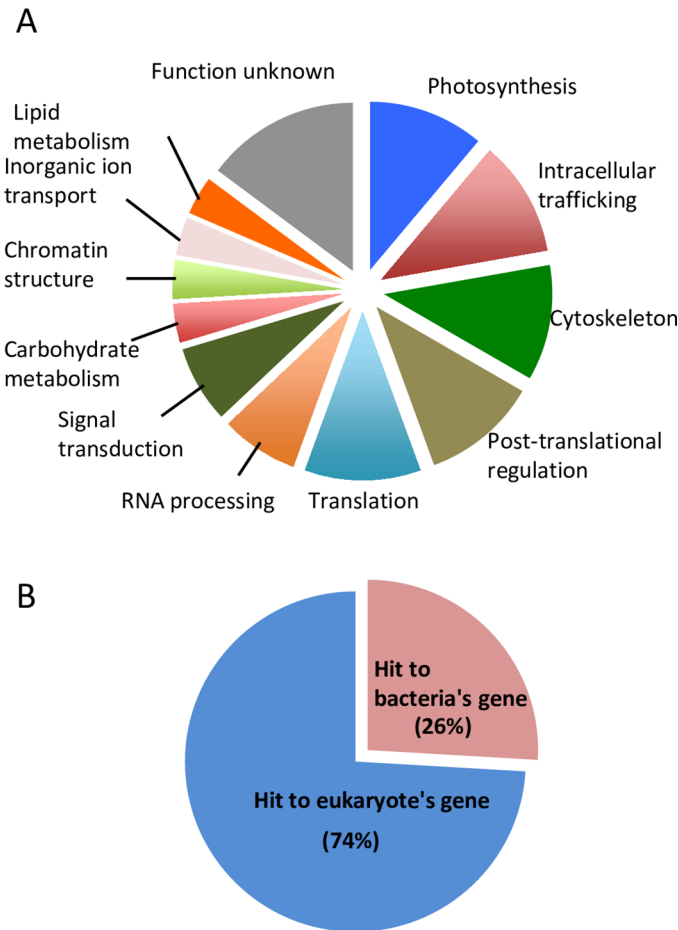
Fig 3. Functional and taxonomical classification of SL *trans*-spliced genes. A. Functional categorization of 27 *trans*-spliced genes of *P. micropora* according to COG and KOG classification [44, 45] with manual editing; a class of photosynthesis related function is added. **B.** Taxonomical classification of 27 SL *trans*-spliced genes based on the BLASTX annotation of the NCBI-nr database.

https://doi.org/10.1371/journal.pone.0200961.g003

and functions (Fig 3A, S1 Table) similar to the cases of other SL *trans*-splicing organisms [4], and we could not detect any bias for specific cellular functions. Next, we examined whether these 27 genes are similar to either of eukaryotic- or prokaryotic-genes by BLASTX (nr). Interestingly, 7 of them (26%) were assigned to bacterial genes (Fig 3B), which include photosynthetic genes from endosymbiotic cyanobacteria [40]. This indicates that the endosymbiotically and horizontally transferred genes could also be targets of SL *trans*-splicing in this organism.

Characterization of the SL RNA genes in *P. micropora*

To obtain insight into the SL RNA gene that donates the SL sequence to the pre-mRNAs, we first performed shotgun genome sequencing of *P. micropora* MYN1 with Illumina Hiseq 2000 (Illumina, San Diego, CA, USA) and obtained *circa* 438 Mb contig sequences in total (data not shown). We then searched these genomic sequences for the SL RNA genes by a query sequence (`TGGATAATCCGGCTTTTCTG-GT`), which contained 20-base conserved SL sequence followed by the 5' splice site motif (GT). Although the genomic contigs were highly fragmented (N50 = 350 bp) because of the large-sized and highly repetitive nature of the *P. micropora* genome, we finally detected 62 candidates for the SL RNA genes. They were classified into two

subtypes according to their distinct intronic sequences, and we named them SL-I and SL-II (S2 Fig and Fig 4).

Their transcription start sites were determined with the aid of cap-trapper cDNA reads, revealing that the TSSs of SL-1 genes were located at the 5′ terminal region (+1 to +5) of the SL exon, while those of SL-II genes were mainly located at the mid region (+19). This suggests that the whole SL exon sequence was transcribed only from SL-I genes but not from SL-II genes (S2 Fig, arrowheads). Cellular RNA levels of the SL-I and SL-II transcripts were compared with those of the 18S rRNA and with the mature and premature forms of calmodulin mRNAs (Fig 4B). The RNA level of SL-I was about 100 times higher than those of SL-II and calmodulin mature mRNA, and was comparable to 1% of 18S rRNA level. This high-level expression of the SL-1 genes, as well as their TSS positions, indicates that the SL-I genes should be the major source of SL sequences of this organism.

To confirm the above results, which were based on NGS analysis of fragmented sequences, we cloned and Sanger sequenced the genomic and cDNA sequences of an SL-I gene, which we named pmSL1 (Fig 4A). pmSL1 is an ncRNA gene with no long ORFs, and its alignment with the other SL-I gene sequences (S2 Fig) showed that the +1 to +130 base region in pmSL1 is highly conserved, implying the functional importance of this region (Fig 4A). This region is predicted to form a tight RNA secondary structure with three stem-loops (Fig 4C). The splice junction is located near the small bulge within the stem-loop I, representing a typical structure of SL RNA genes [1, 4]. SL RNAs generally encode an Sm-binding site required for SL RNP formation [12, 47, 48]; it is usually located at the single strand region flanking the stem-loops, and its consensus sequence is $AAU_{3-6}GG/U$ in trypanosome and nematode [1, 47, 49]. However, we could not find any such typical Sm binding sequence in the conserved region of pmSL1. In *P. micropora*, the Sm binding sequence may have diverged from the canonical ones.

To know the 3′ end formation of the pmSL1 RNA, we searched downstream of the above conserved region for the pyrimidine stretch sequence ($\geq$ 6b) containing poly(T) that is used for the transcription termination signal of SL RNA genes in protists [26, 50, 51]. However, we could not find any such pyrimidine stretch; instead, we found a canonical polyadenylation signal, AATAAA, 70 bases downstream of the conserved region (Fig 4A). 3′ RACE analysis showed that pmSL1 RNA was polyadenylated 14 bases downstream of this putative poly(A) signal (Fig 4A, asterisks). In *P. micropora*, transcriptional termination of SL RNA genes is likely to occur in a poly(A) signal-dependent manner, which is quite different from that of other SL *trans*-splicing organisms.

Discussion

This study showed that SL *trans*-splicing occurs in a rhizarian thecate amoeba, the *P. micropora* MYN1 strain. Its SL is short (33 bp), contains the UUU triplet, and has G at the 3′ end (Fig 1). This SL *trans*-splicing occurs for genes having diverse cellular functions (Fig 3). In addition, SL RNA can form tight secondary structures (Fig 4). These characteristics of *P. micropora's* SL *trans*-splicing represent general features shared by other SL *trans*-splicing organisms known thus far [1, 4]. We found 27 protein-coding genes, most likely SL *trans*-spliced (Fig 1 and S1 Table). However, this number should be the tip of the iceberg, because the number of cap-trapper cDNA reads available in this study was far less than the number covering the total gene (data not shown). To depict the overview of the SL trans-splicing in this organism, more comprehensive analyses of its genome and transcriptome are needed.

Although the *P. micropora* SL sequence (33b) shares common structural features with those of other SL *trans*-splicing organisms, it does not show sequence conservation except that of *P. chromatophora*: a member of the same photosynthetic *Paulinella* group. This is not surprising,

Fig 4. Characteristics of SL RNA genes in *P. micropora*. A. ClustalW alignment of SL-I RNA gene candidates. pmSL1 represents the Sanger sequence of a PCR-amplified clone, and the others are *de novo* assembled contig sequences from *P. micropora* genomic shotgun sequencing. Transcription start sites (arrowheads) of pmSL1 were determined by 5′ end mapping of cap-trapper cDNA reads, and the cleavage sites (asterisks) were determined by 3′ RACE analysis. Canonical poly(A) signals (AATAAA) are indicated by red boxes. The regions forming the stem-loop structure in pmSL1 are represented by dashed lines. **B.** Relative levels of SL-I RNA, SL-II RNA, and calmodulin (CALM) mRNA to 18S rRNA. Bulk expression levels of SL-I and SL-II genes (S2 Fig) were analyzed by real-time RT-PCR using conserved sequences of the respective SL RNA gene group. **C.** Secondary structure of the conserved region of pmSL1 RNA predicted by Mfold [46]. The shaded area indicates the candidate of Sm binding site, when the consensus rule is loosened to pyrimidine rich sequence sandwiched by adenosine and guanosine.

https://doi.org/10.1371/journal.pone.0200961.g004

because the *Paulinella* species are phylogenetically distant from other SL *trans*-splicing organisms at the kingdom level [52], and its SL sequence is not conserved at the interphylum level [4]. In nematodes, the SL sequence is involved in the transcriptional regulation of SL RNA genes [53] as well as the translational regulation of *trans*-spliced messages [54]. If this is also the case for *P. micropora*, its unique SL sequence may mediate similar functions.

SL *trans*-splicing is observed in two *Paulinella* species, *P. micropora* MYN1 (this study) and *P. chromatophora* CCAC0185 [27], which are thought to have diverged about 45.7–64.7 million years ago [37, 42]. Therefore, the SL *trans*-splicing ability should have been acquired by the *Paulinella* lineage earlier than their divergence. Notably, SL *trans*-splicing has not been found in other rhizarian organisms: *Bigelowiella natans* [55], *Plasmodiophora brassicae* [56], and *Reticulomyxa filose* [57]. *B. natans* and *P. brassicae* belong to cercozoa, the same phylum as the *Paulinella* species. Therefore, SL *trans*-splicing should have been acquired uniquely in the *Paulinella* lineage, or multiple independent losses might have occurred in other rhizarian lineages.

In *P. micropora*, we found two subtypes of SL RNA genes, namely, SL-I and SL-II, with distinct intronic sequences; we detected 43 SL-I gene copies and 14 SL-II gene copies. Multiple gene copies are general features of SL RNA genes. Because available genome data for *P. micropora* were rather incomplete in this study, we expect many more gene copies to be present for SL-I and SL-II. From their expression levels, we expect SL-I genes to be the major source of SL RNAs in this organism, while we do not know whether SL-II genes have specific functions. (Fig 4B).

In this study, we cloned and analyzed an SL-I gene copy, pmSL1, which revealed two intriguing features. One lies in its 3′ end formation. SL RNAs in metazoan, kinetoplastid, and dinoflagellates are usually transcribed as poly(A)-less RNA [2, 3, 25, 58]. A few SL RNAs are exceptionally polyadenylated in dinoflagellates [25] and kinetoplastida [59, 60], with the poly(A) addition occurring at processing sites upstream of the poly(T) tract [50, 59, 60]. In this respect, pmSL1 in *P. micropora* is unique in utilizing a canonical polyadenylation signal (AAUAAA) for the addition of poly(A) (Fig 4A). In Fig 4A, 4 of 12 SL-I genes possess the canonical poly(A) signal downstream of the 130-base conserved region, suggesting that at least some portion of the SL-I RNAs are polyadenylated. Poly(T) tract and 3′-box sequence (GTTTAAAACAAGC), found at the transcriptional terminator regions of SL RNA genes in nematode [61], are not found in SL-I RNA genes (Fig 4A). Transcriptional termination and 3′ end formation of SL RNA genes in *P. micropora* occurs differently from that of other SL *trans*-splicing organisms.

Another intriguing feature of pmSL1 is the absence of a typical Sm binding sequence (AAU$_{3-6}$GG/U) in the 130-base conserved region. The Sm binding sequence is essential in forming SL RNP and is usually conserved in SL RNAs across the phyla. In relation to this, the Sm binding sequence may have diverged according to Sm protein variants [1, 19, 62]; for example, the Sm binding site of SL RNA (AACUCUCUCCUAUCCCUCUCG) in a tunicate, *Oikopleura dioica*, is far diverged from the canonical sequence [29]. Therefore, it may be possible that *P. micropora* has a highly diverged version of the Sm binding sequence. If the consensus sequence rule is relaxed to pyrimidine rich sequence sandwiched by adenine and guanine on the single strand region of SL RNA, a candidate of the Sm binding sequence (ACCUCUCCCU UCUCUUCG) similar to that of *O. dioica* is found in the pmSL1 (Fig 4). Further investigation is necessary to elucidate how *P. micropora's* SL RNA is involved in SL RNP formation for *trans*-splicing.

This study showed that in *P. micropora*, the nuclear genes derived from EGT and HGT are included in the targets of SL *trans*-splicing (Fig 3B). Because this organism is undergoing endosymbiotic evolution [36, 39, 40], it is intriguing to consider whether SL *trans*-splicing can

contribute to its evolutionary process. If a given organism undergoes frequent EGT and HGT, its intrinsic genes in the genome might suffer from accidental insertion of the alien gene into the 5' UTR, which disturbs the translation of the original coding sequence (CDS). This pitfall might be circumvented by SL *trans*-splicing, which directly adds the 5′ cap plus translationally favorable UTR sequence to the original CDS to restore its translatability. From the RNA-seq data of *P. micropora*, we found a case that appears to fit our prediction (S4 Fig, S3 Table). In addition, rotifer, an SL *trans*-splicing organism, contains an unusually high number of foreign genes that are acquired horizontally [63, 64]. Therefore, SL *trans*-splicing might be involved in facilitating HGT/EGT. Interestingly, in dinoflagellate, SL *trans*-splicing should have contributed to the amplification and maintenance of the retrotransposed gene, resulting in massive gene birth during adaptive evolution [65]. The facilitative role of SL *trans*-splicing in functional gene transfer and eukaryotic genome evolution deserves further investigation.

In this study, we characterized the SL *trans*-splicing of a photosynthetic thecate amoeba, *Paulinella micropora*, and implied its possible role in HGT/EGT. Further comprehensive analysis of the genome and transcriptome of this organism will provide further insight into the possible contribution of SL *trans*-splicing to endosymbiotic evolution.

Materials and methods

Cell culture, RNA- and DNA-extraction

The *P. micropora* MYN1 strain was renamed from *P. chromatophora* MYN1 [41] on the basis of the morphological trait of *P. micropora* [42] and phylogenetic analysis (S3 Fig). The *P. micropora* MYN1 strain, which has been deposited at the National Institute for the Environmental Sciences, Tsukuba as NIES-4060, was cultured according to Nomura et al. [41]. The cells were harvested at low speed centrifuge (500 g × 2 min) at 4˚C. Total RNAs were extracted using Trizol (Thermo Fisher Scientific) and the total genome was isolated by DNeasy Plant Mini Kit (Qiagen, Hilden, Germany).

Cap-trapper cDNA library analysis

The *P. micropora* cap-trapper cDNA library was constructed according to Carninci and Haya-shizaki [43] with slight modifications. 15 μg total RNA sample was used as starting material. Reverse transcription was carried out using the adapter primer plus 15-base random sequence [66]. After cap-trapping, linker ligation and second strand cDNA synthesis were performed according to Shiraki et al. [67]. cDNA products (150–350 bp) were purified by agarose gel electrophoresis and subjected to PCR (32–35 cycles) using linker- and adapter-specific primers. After PCR, 150–350 bp fraction was purified and subjected to the preparation of sequencing libraries with Ion Plus Fragment Library Kit (Thermo Fisher Scientific). Sequencing analyses were performed with Ion PGM^TM Sequencer (Thermo Fisher Scientific) (DDBJ Accession No. DRA004751). To determine the TSSs of the calmodulin gene, calmodulin cDNAs were amplified with the linker and calmodulin gene specific primers from the cap-trapper cDNA pool. The linker, adapter, and PCR primer sequences used in this study are described in S2 Table.

Searching for SL *trans*-spliced gene sequences using cap-trapper cDNA library

Linker- and adapter-sequences of the cap-trapper library were trimmed by Cutadapt software [68], and then short reads (<30b) and reads hitting to rRNA and *P. chromatophora's* genes [36] by BLASTN (e-value<1e-10) were discarded. During SL sequence screening, the reads hitting N-terminal 50 amino acid residues of eukaryotic genes in the NCBI-nr database were

collected (BLASTX, e-value<1e-5), and then assembled into contigs by CLC Genomic Work-bench with default settings (CLC bio, Tokyo, Japan). The contigs' information and BLASTX results were used to group the cap-trapper reads. The common sequences at the 5′ end of the reads were searched using BLASTN algorithm.

Genome sequencing, gene structure, and expression analysis

Shotgun genomic sequencing of the *P. micropora* genome was performed with Illumina Hiseq2000 (Illumina) (DDBJ Accession No. DRA004743) and assembled by Velvet [69]. The library construction, sequencing, and genome assembly were provided as a customary service by Eurofins MWG Operon LLC (Tokyo, Japan). PCR, RT-PCR, and Sanger sequencing analyses were performed by gene specific primers listed in S2 Table. Gene and cDNA sequences analyzed by Sanger sequencing method are available in DDBJ databank (LC383945, LC383946, LC384061–LC384070). Genome contig sequences associated with SL RNA genes are shown in S3 Table. For gene expression analysis, cDNAs were synthesized with random nonamers and ReverTra Ace (Toyobo Bio-Technology Co., Ltd, Osaka, Japan), then subjected to real-time with Thunderbird SYBR qPCR mix (Tobyobo Bio-Technology, Co., Ltd, Osaka, Japan) and Eco Real-Time PCR System (Illumina). In a 3′ RACE experiment, cDNAs were synthesized using an anchored Oligo $(dT)_{16}$ primer from poly(A)+ RNA, purified with Dyna-beads® mRNA DIRECT™ Kit (Thermo Fisher Scientific), and amplified with anchor primer and SL RNA specific primer by PCR. SL RNA structure was predicted by Mfold [46] (also see http://www.bioinfo.rpi.edu/applications/mfold).

Supporting information

S1 Table. COG/KOG classification of SL-containing cDNA reads.
(XLSX)

S2 Table. Primers and oligonucleotides.
(XLSX)

S3 Table. Sequences analyzed in Fig 4, S2 Fig and S4 Fig.
(XLSX)

S1 Fig. Mapping of the cap-trapper cDNA tag-reads on actin and polyubiquitin genes. 5' ends of the tag-reads were mapped by BLASTN (identity ≥ 95%, alignment length/ the read length> 0.9).
(PDF)

S2 Fig. SL gene candidates in *P. micropora* genome. SL gene candidates detected in the *P. micropora* genome contigs by BLASTN analysis using 22bp query sequence, which consists of 20bp SL-conserved sequence (TGGATAATCCGGCTTTTCTG) and 5' splicing motif (GT). The sequences were aligned by MAFFT (ver.7)# and were grouped according to the intron sequences. Exon- and intron-regions of SL-I genes are indicated by solid and dashed bold lines above the sequences. Transcription start sites and their nucleotides, where 5' end of cap trapper cDNA reads are mapped, are shown by arrowheads and red color, respectively. Asterisk means the end of the contig sequence. # Katoh *et al.* (2002) *Nucleic Acids Res* 30, 3059–3066.
(PDF)

S3 Fig. Maximum likelihood tree of 16S rRNA gene sequences of photosynthetic *Paulinella* species. The sequences are analyzed using Kimura-2 +G model by MEGA7 [1] with 1000 boot-strap replication. IDs in parentheses are Genbank- and DDBJ-accessions. *P. micropora* FK01 (*) and MYN1 (**), previously reported as *P. chromatophora* strains [2,3], were renamed in

Lhee et al. [4] and in this study, respectively.

[1] Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 2016; 33: 1870–4.

[2] Yoon HS, Nakayama T, Reyes-Prieto A, Andersen RA, Boo SM, Ishida K, Bhattacharya D. A single origin of the photosynthetic organelle in different *Paulinella* lineages. BMC Evol. Biol. 2009; 9: 98.

[3] Nomura M, Nakayama T, Ishida K. Detailed process of shell construction in the photosynthetic testate amoeba *Paulinella chromatophora* (euglyphid, Rhizaria). J. Eukaryot. Microbiol. 2014; 61: 317–21.

[4] Lhee D, Yang EC, Kim JI, Nakayama T, Zuccarello G, Andersen RA, Yoon HS. Diversity of the Photosynthetic *Paulinella* Species, with the Description of *Paulinella micropora* sp. nov. and the Chromatophore Genome Sequence for strain KR01. Protist. 2017; 168: 155–70.
(PDF)

S4 Fig. A case of SL *trans*-spliced mRNA rescued from the disturbance by the retrotransposon insertion just upstream of the coding region. A. Schematic models of the premature and SL *trans*-sliced forms of the transcripts encoding a serine/threonine kinase. These RNA sequences were searched from the RNA-seq data of *P. micropora* obtained by Illumina HiSeq 4000. **B.** Nucleotide sequences covering the *trans*-splicing site and translation initiation context of the transcripts modeled in **A**, with the 3' splicing acceptor motif (AG), termination codon, and initiation codon shown by red characters.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: Mitsuhiro Matsuo, Junichi Obokata.

Data curation: Mitsuhiro Matsuo, Atsushi Katahata, Soichirou Satoh, Motomichi Matsuzaki, Mami Nomura, Ken-ichiro Ishida, Yuji Inagaki.

Formal analysis: Mitsuhiro Matsuo, Atsushi Katahata.

Funding acquisition: Mitsuhiro Matsuo, Yuji Inagaki, Junichi Obokata.

Investigation: Mitsuhiro Matsuo, Atsushi Katahata, Soichirou Satoh, Motomichi Matsuzaki.

Project administration: Junichi Obokata.

Resources: Mami Nomura, Ken-ichiro Ishida.

Supervision: Junichi Obokata.

Writing – original draft: Mitsuhiro Matsuo.

Writing – review & editing: Junichi Obokata.

References

1. Lasda EL, Blumenthal T. Trans-splicing. Wiley Interdisciplinary Reviews: RNA. 2011; 2: 417–34. https://doi.org/10.1002/wrna.71 PMID: 21957027

2. Murphy WJ, Watkins KP, Agabian N. Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: Evidence for Trans splicing. Cell. 1986; 47:517–25. PMID: 3779835

3. Sutton RE, Boothroyd JC. Evidence for Trans splicing in trypanosomes. Cell. 1986; 47: 527–35. PMID: 3022935

4. Bitar M, Boroni M, Macedo AM, Machado CR, Franco GR. The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. Frontiers in genetics. 2013; 4: 199. https://doi.org/10.3389/fgene.2013.00199 PMID: 24130571

5. Bruzik JP, Maniatis T. Spliced leader RNAs from lower eukaryotes are trans-spliced in mammalian cells. Nature. 1992; 360: 692–95. https://doi.org/10.1038/360692a0 PMID: 1465136

6. Tschudi C, Ullu E. Destruction of U2, U4, or U6 small nuclear RNA blocks Trans splicing in trypanosome cells. Cell. 1990; 61: 459–66. PMID: 1692263

7. Hannon GJ, Maroney PA, Nilsen TW. U small nuclear ribonucleoprotein requirements for nematode cis- and trans-splicing in vitro. J. Biol. Chem. 1991; 266: 22792–5. PMID: 1835972

8. Maroney PA, Yu YT, Jankowska M, Nilsen TW. Direct analysis of nematode cis- and trans-spliceo-somes: a functional role for U5 snRNA in spliced leader addition trans-splicing and the identification of novel Sm snRNPs. RNA. 1996; 2: 735–45. PMID: 8752084

9. Liang X-h, Haritan A, Uliel S, Michaeli S. trans and cis splicing in trypanosomatids: mechanism, factors, and regulation. Eukaryot. Cell. 2003; 2: 830–40. https://doi.org/10.1128/EC.2.5.830-840.2003 PMID: 14555465

10. Luz Ambrósio D, Lee JH, Panigrahi AK, Nguyen TN, Cicarelli RMB, Günzl A. Spliceosomal proteomics in Trypanosoma brucei reveal new RNA splicing factors. Eukaryot. Cell. 2009; 8: 990–1000. https://doi.org/10.1128/EC.00075-09 PMID: 19429779

11. Tkacz ID, Gupta SK, Volkov V, Romano M, Haham T, Tulinski P, et al. Analysis of spliceosomal proteins in Trypanosomatids reveals novel functions in mRNA processing. J. Biol. Chem. 2010; 285: 27982–99. https://doi.org/10.1074/jbc.M109.095349 PMID: 20592024

12. Thomas JD, Conrad RC, Blumenthal T. The C. elegans trans-spliced leader RNA is bound to Sm and has a trimethylguanosine cap. Cell. 1988; 54: 533–9. PMID: 3401926

13. Ferguson KC, Heid PJ, Rothman JH. The SL1 trans-spliced leader RNA performs an essential embry-onic function in Caenorhabditis elegans that can also be supplied by SL2 RNA. Genes Dev. 1996; 10: 1543–56. PMID: 8666237

14. Lustig Y, Sheiner L, Vagima Y, Goldshmidt H, Das A, Bellofatto V, et al. Spliced-leader RNA silencing: a novel stress-induced mechanism in Trypanosoma brucei. EMBO rep. 2007; 8: 408–13. https://doi.org/10.1038/sj.embor.7400930 PMID: 17347669

15. Johnson PJ, Kooter JM, Borst P. Inactivation of transcription by UV irradiation of T. brucei provides evidence for a multicistronic transcription unit including a VSG gene. Cell. 1987; 51: 273–81. PMID: 3664637

16. Muhich ML, Boothroyd JC. Polycistronic transcripts in trypanosomes and their accumulation during heat shock: evidence for a precursor role in mRNA synthesis. Mol. Cell. Biol. 1988; 8: 3837–46. https://doi.org/10.1128/MCB.8.9.3837 PMID: 3221866

17. Tschudi C, Ullu E. Polygene transcripts are precursors to calmodulin mRNAs in trypanosomes. EMBO J. 1988; 7: 455–63. PMID: 3366120

18. Spieth J, Brooke G, Kuersten S, Lea K, Blumenthal T. Operons in C. elegans: Polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. Cell. 1993; 73:521–32. PMID: 8098272

19. Davis RE. Spliced leader RNA trans-splicing in metazoa. Parasitol. Today. 1996; 12: 33–40. PMID: 15275306

20. Zeiner GM, Sturm NR, Campbell DA. The Leishmania tarentolae spliced leader contains determinants for association with polysomes. J. Biol. Chem. 2003; 278: 38269–75. https://doi.org/10.1074/jbc.M304295200 PMID: 12878606

21. Hastings KEM. SL trans-splicing: easy come or easy go? Trends Genet. 2005; 21: 240–7. https://doi.org/10.1016/j.tig.2005.02.005 PMID: 15797620

22. Cheng G, Cohen L, Ndegwa D, Davis RE. The flatworm spliced leader 3'-terminal AUG as a translation initiator methionine. J. Biol. Chem. 2006; 281: 733–43. https://doi.org/10.1074/jbc.M506963200 PMID: 16230357

23. Douris V, Telford MJ, Averof M. Evidence for Multiple Independent Origins of trans-Splicing in Metazoa. Mol. Biol. Evol. 2010; 27: 684–93. https://doi.org/10.1093/molbev/msp286 PMID: 19942614

24. Tessier LH, Keller M, Chan RL, Fournier R, Weil JH, Imbault P. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in Euglena. EMBO J. 1991; 10: 2621–5. PMID: 1868836

25. Zhang H, Hou Y, Miranda L, Campbell DA, Sturm NR, Gaasterland T, et al. Spliced leader RNA trans-splicing in dinoflagellates. Proc. Natl. Acad. Sci. U.S.A. 2007; 104: 4618–23. https://doi.org/10.1073/pnas.0700258104 PMID: 17360573

26. Zhang H, Campbell DA, Sturm NR, Dungan CF, Lin S. Spliced leader RNAs, mitochondrial gene frame-shifts and multi-protein phylogeny expand support for the genus Perkinsus as a unique group of alveolates. PLoS ONE. 2011; 6(5): e19933. https://doi.org/10.1371/journal.pone.0019933 PMID: 21629701

27. Nowack ECM, Price DC, Bhattacharya D, Singer A, Melkonian M, Grossman AR. Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of Paulinella chromatophora. Proc. Natl. Acad. Sci. U.S.A. 2016; 113: 12214–9. https://doi.org/10.1073/pnas.1608016113 PMID: 27791007

28. Vandenberghe AE, Meedel TH, Hastings KE. mRNA 5'-leader trans-splicing in the chordates. Genes Dev. 2001; 15:294–303. https://doi.org/10.1101/gad.865401 PMID: 11159910

29. Ganot P, Kallesøe T, Reinhardt R, Chourrout D, Thompson EM. Spliced-leader RNA trans splicing in a chordate, Oikopleura dioica, with a compact genome. Mol. Cell. Biol. 2004; 24: 7795–805. https://doi.org/10.1128/MCB.24.17.7795-7805.2004 PMID: 15314184

30. Yang F, Xu D, Zhuang Y, Yi X, Huang Y, Chen H, et al. Spliced leader RNA trans-splicing discovered in copepods. Sci Rep. 2015; 5: 17411. https://doi.org/10.1038/srep17411 PMID: 26621068

31. Stover NA, Steele RE. Trans-spliced leader addition to mRNAs in a cnidarian. Proc. Natl. Acad. Sci. U.S.A. 2001; 98: 5693–8. https://doi.org/10.1073/pnas.101049998 PMID: 11331766

32. Derelle R, Momose T, Manuel M, Da Silva C, Wincker P, Houliston E. Convergent origins and rapid evolution of spliced leader trans-splicing in metazoa: insights from the ctenophora and hydrozoa. RNA. 2010; 16: 696–707. https://doi.org/10.1261/rna.1975210 PMID: 20142326

33. Hannon GJ, Maroney PA, Yu YT, Hannon GE, Nilsen TW. Interaction of U6 snRNA with a sequence required for function of the nematode SL RNA in trans-splicing. Science. 1992; 258: 1775–80. PMID: 1465612

34. Marin B, Nowack ECM, Melkonian M. A plastid in the making: evidence for a second primary endosymbiosis. Protist. 2005; 156: 425–32. https://doi.org/10.1016/j.protis.2005.09.001 PMID: 16310747

35. Yoon HS, Reyes-Prieto A, Melkonian M, Bhattacharya D. Minimal plastid genome evolution in the Paulinella endosymbiont. Curr. Biol. 2006; 16: R670–2. https://doi.org/10.1016/j.cub.2006.08.018 PMID: 16950085

36. Nowack ECM, Melkonian M, Glöckner G. Chromatophore genome sequence of Paulinella sheds light on acquisition of photosynthesis by eukaryotes. Curr. Biol. 2008; 18: 410–8. https://doi.org/10.1016/j.cub.2008.02.051 PMID: 18356055

37. Delaye L, Valadez-Cano C, Pérez-Zamorano B. How Really Ancient Is Paulinella Chromatophora? PLOS Currents Tree of Life. 2016. Edition 1. https://doi.org/10.1371/currents.tol.e68a099364bb1a1e129a17b4e06b0c6b PMID: 28515968

38. Sánchez-Baracaldo P, Raven JA, Pisani D, Knoll AH. Early photosynthetic eukaryotes inhabited low-salinity habitats. Proc. Natl. Acad. Sci. U.S.A. 2017; 114: E7737–E45. https://doi.org/10.1073/pnas.1620089114 PMID: 28808007

39. Nakayama T, Ishida K-i. Another acquisition of a primary photosynthetic organelle is underway in Paulinella chromatophora. Curr. Biol. 2009; 19: R284–5. https://doi.org/10.1016/j.cub.2009.02.043 PMID: 19368869

40. Nowack E, Vogel H, Groth M, Grossman A, Melkonian M, Glöckner G. Endosymbiotic gene transfer and transcriptional regulation of transferred genes in Paulinella chromatophora. Mol. biol. Evol. 2011; 28: 407–22. https://doi.org/10.1093/molbev/msq209 PMID: 20702568

41. Nomura M, Nakayama T, Ishida K-i. Detailed process of shell construction in the photosynthetic testate amoeba Paulinella chromatophora (euglyphid, Rhizaria). J. Eukaryot. Microbiol. 2014; 61: 317–21. https://doi.org/10.1111/jeu.12102 PMID: 24444111

42. Lhee D, Yang EC, kim JI, Nakayama T, Zuccarello G, Andersen RA, et al. Diversity of the Photosynthetic Paulinella Species, with the Description of Paulinella micropora sp. nov. and the Chromatophore Genome Sequence for strain KR01. Protist. 2017; 168: 155–70. https://doi.org/10.1016/j.protis.2017.01.003 PMID: 28262587

43. Carninci P, Hayashizaki Y. High-efficiency full-length cDNA cloning. Methods Enzymol. 1999; 303: 19–44. PMID: 10349636

44. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC bioinformatics. 2003; 4(1): 1. https://doi.org/10.1186/1471-2105-4-41 PMID: 12969510

45. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015; 43(Database issue): D261–9. https://doi.org/10.1093/nar/gku1223 PMID: 25428365

46. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003; 31: 3406–15. PMID: 12824337

47. Denker JA, Maroney PA, Yu YT, Kanost RA, Nilsen TW. Multiple requirements for nematode spliced leader RNP function in trans-splicing. RNA. 1996; 2: 746–55. PMID: 8752085

48. Mandelboim M, Barth S, Biton M, Liang X-h, Michaeli S. Silencing of Sm proteins in Trypanosoma brucei by RNA interference captured a novel cytoplasmic intermediate in spliced leader RNA biogenesis. J. Biol. Chem. 2003; 278: 51469–78. https://doi.org/10.1074/jbc.M308997200 PMID: 14532264

49. Gibson W, Bingle L, Blendeman W, Brown J, Wood J, Stevens J. Structure and sequence variation of the trypanosome spliced leader transcript. Mol. Biochem. Parasitol. 2000; 107: 269–77. PMID: 10779603

50. Sturm NR, Yu MC, Campbell DA. Transcription termination and 3'-End processing of the spliced leader RNA in kinetoplastids. Mol. Cell. Biol. 1999; 19: 1595–604. https://doi.org/10.1128/MCB.19.2.1595 PMID: 9891092

51. Zhang H, Campbell DA, Sturm NR, Lin S. Dinoflagellate spliced leader RNA genes display a variety of sequences and genomic arrangements. Mol. Biol. Evol. 2009; 26: 1757–71. https://doi.org/10.1093/molbev/msp083 PMID: 19387009

52. Burki F. The eukaryotic tree of life from a global phylogenomic perspective. Cold Spring Harbor Perspectives in Biology. 2014; 6(5): a016147–a. https://doi.org/10.1101/cshperspect.a016147 PMID: 24789819

53. Hannon GJ, Maroney PA, Ayers DG, Shambaugh JD, Nilsen TW. Transcription of a nematode trans-spliced leader RNA requires internal elements for both initiation and 3' end-formation. EMBO J. 1990; 9: 1915–21. PMID: 2347310

54. Lall S, Friedman CC, Jankowska-Anyszka M, Stepinski J, Darzynkiewicz E, Davis RE. 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. https://doi.org/10.1074/jbc.M407475200 PMID: 15322127

55. Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, et al. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature. 2012; 492: 59–65. https://doi.org/10.1038/nature11681 PMID: 23201678

56. Schwelm A, Fogelqvist J, Knaust A, Jülke S, Lilja T, Bonilla-Rosso G, et al. The Plasmodiophora brassicae genome reveals insights in its life cycle and ancestry of chitin synthases. Sci Rep. 2015; 5: 11153. https://doi.org/10.1038/srep11153 PMID: 26084520

57. Glöckner G, Hülsmann N, Schleicher M, Noegel AA, Eichinger L, Gallinger C, et al. The genome of the foraminiferan Reticulomyxa filosa. Curr. Biol. 2014; 24: 11–8. https://doi.org/10.1016/j.cub.2013.11.027 PMID: 24332546

58. Krause M, Hirsh D. A trans-spliced leader sequence on actin mRNA in C. elegans. Cell. 1987; 49: 753–61. PMID: 3581169

59. Pellé R, Murphy NB. Stage-specific differential polyadenylation of mini-exon derived RNA in African trypanosomes. Mol. Biochem. Parasitol. 1993; 59: 277–86. PMID: 7688101

60. Lamontagne J, Papadopoulou B. Developmental regulation of spliced leader RNA gene in Leishmania donovani amastigotes is mediated by specific polyadenylation. J. Biol. Chem. 1999; 274: 6602–9. https://doi.org/10.1074/jbc.274.10.6602 PMID: 10037755

61. Nilsen TW, Shambaugh J, Denker J, Chubb G, Faser C, Putnam L, et al. Characterization and expression of a spliced leader RNA in the parasitic nematode Ascaris lumbricoides var. suum. Mol. Cell. Biol. 1989; 9: 3543–7. https://doi.org/10.1128/MCB.9.8.3543 PMID: 2796996

62. Wang P, Palfi Z, Preusser C, Lücke S, Lane WS, Kambach C, et al. Sm core variation in spliceosomal small nuclear ribonucleoproteins from Trypanosoma brucei. EMBO J. 2006; 25: 4513–23. https://doi.org/10.1038/sj.emboj.7601328 PMID: 16977313

63. Pouchkina-Stantcheva NN, Tunnacliffe A. Spliced Leader RNA–Mediated trans-Splicing in Phylum Rotifera. Mol. Biol. Evol. 2005; 22: 1482–9. https://doi.org/10.1093/molbev/msi139 PMID: 15788744

64. Gladyshev EA, Meselson M, Arkhipova IR. Massive horizontal gene transfer in bdelloid rotifers. Science. 2008; 320: 1210–3. https://doi.org/10.1126/science.1156407 PMID: 18511688

65. Song B, Morse D, Song Y, Fu Y, Lin X, Wang W, et al. Comparative Genomics Reveals Two Major Bouts of Gene Retroposition Coinciding with Crucial Periods of Symbiodinium Evolution. Genome biology and evolution. 2017; 9: 2037–47. https://doi.org/10.1093/gbe/evx144 PMID: 28903461

66. Stangegaard M, Dufva IH, Dufva M. Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA. BioTechniques. 2006; 40: 649–57. https://doi.org/10.2144/000112153 PMID: 16708763

67. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. U.S.A. 2003; 100: 15776–81. https://doi.org/10.1073/pnas.2136655100 PMID: 14663149

68. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal. 2011; 17(1): 10–2. https://doi.org/10.14806/ej.17.1.200

69. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18: 821–9. https://doi.org/10.1101/gr.074492.107 PMID: 18349386