Evidence for Multiple Independent Origins of *trans*-Splicing in Metazoa

Vassilis Douris,†,[1] Maximilian J. Telford,†,[2] and Michalis Averof*,[1]

[1]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology Hellas, Iraklio, Crete, Greece

[2]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

†These authors contributed equally to this work.

***Corresponding author:** E-mail: averof@imbb.forth.gr.

Associate editor: Hervé Philippe

Research article

Abstract

In contrast to conventional splicing, which joins exons from a single primary transcript, *trans*-splicing links stretches of RNA from separate transcripts, derived from distinct regions of the genome. Spliced leader (SL) *trans*-splicing is particularly well known in trypanosomes, nematodes, and flatworms, where it provides messenger RNAs with a leader sequence and cap that allow them to be translated efficiently. One of the largest puzzles regarding SL *trans*-splicing is its evolutionary origin. Until now SL *trans*-splicing has been found in a small and disparate set of organisms (including trypanosomes, dinoflagellates, cnidarians, rotifers, nematodes, flatworms, and urochordates) but not in most other eukaryotic lineages, including well-studied groups such as fungi, plants, arthropods, and vertebrates. This patchy distribution could either suggest that *trans*-splicing was present in early eukaryotes/metazoans and subsequently lost in multiple lineages or that it evolved several times independently. Starting from the serendipitous discovery of SL *trans*-splicing in an arthropod, we undertook a comprehensive survey of this process in the animal kingdom. By surveying expressed sequence tag data from more than 70 metazoan species, we show that SL *trans*-splicing also occurs in at least two groups of arthropods (amphipod and copepod crustaceans), in ctenophores, and in hexactinellid sponges. However, we find no evidence for SL *trans*-splicing in other groups of arthropods and sponges or in 15 other phyla that we have surveyed. Although the presence of SL *trans*-splicing in hydrozoan cnidarians, hexactinellid sponges, and ctenophores might suggest that it was present at the base of the Metazoa, the patchy distribution that is evident at higher resolution suggests that SL *trans*-splicing has evolved repeatedly among metazoan lineages. In agreement with this scenario, we discuss evidence that SL precursor RNAs can readily evolve from ubiquitous small nuclear RNAs that are used for conventional splicing.

Key words: *trans*-splicing, spliced leader, snRNA, animals, origin, evolution.

Introduction

Splicing is one of the key processes that control the flow of genetic information from DNA to protein and provides opportunities for the evolution of new genes through exon shuffling. Besides conventional (*cis*-) splicing, which joins exons located in the same primary transcript, a small number of organisms are known to possess *trans*-splicing—the ability to link specific sequences present on separate RNA molecules (Bonen 1993; Maniatis and Tasic 2002). By joining sequences that have been transcribed from different genomic locations, *trans*-splicing has the potential to reshape the composition of messenger RNAs (mRNAs) in more radical ways than *cis*-splicing, yet the phenomenon of *trans*-splicing is still not well explored in terms of mechanism, functions, and evolutionary dynamics.

The most widely known form of *trans*-splicing involves the addition of a common spliced leader (SL) sequence to a number of unrelated mRNAs, providing each with a new 5′ cap and leader sequence (Van der Ploeg 1986; Hastings 2005). In this form of *trans*-splicing, the SL sequences come from small transcripts that resemble small nuclear RNAs (snRNAs), carrying an "Sm motif" that binds Sm proteins and leads to the assembly of small nuclear ribonucleoprotein (snRNP) particles (Bruzik et al. 1988; Maroney et al. 1990; Denker et al. 1996). The recruitment of Sm proteins also leads to modification of the 5′ cap structure of SL RNAs from monomethylguanosine (MMG; characteristic of most mRNAs) to trimethylguanosine (TMG; characteristic of snRNAs; Van Doren and Hirsh 1990; Will and Luhrmann 2001). The *trans*-splicing reaction involves most of the core *cis*-spliceosomal RNAs, except U1 snRNP, and uses splice donor and splice acceptor sites that are very similar to those used for *cis*-splicing (Bruzik and Steitz 1990; Hannon et al. 1991; Bruzik and Maniatis 1992; Conrad et al. 1993). These similarities suggest that the mechanisms of *cis*- and *trans*-splicing are closely related.

The functions of SL *trans*-splicing are not fully understood, but the process is generally thought to impinge on the ability of a message to be translated. In some cases, *trans*-splicing is known to provide a 5′ cap or an AUG start codon that are essential for translation: In trypanosomes, it provides a 5′ cap to mRNAs transcribed by RNA polymerase I, which would otherwise be uncapped (Lee and Van der Ploeg 1997); in several organisms, it serves to resolve multicistronic transcripts into capped monocistronic mRNAs (Blumenthal 2005; Satou et al. 2006;

© The Author 2009. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org

Marletaz et al. 2008); and in some flatworm mRNAs, the SL sequence provides the AUG codon for initiating translation (Cheng et al. 2006). In other cases, addition of the SL and TMG cap has been suggested to improve translational efficiency (Maroney et al. 1995; Zeiner et al. 2003; Lall et al. 2004) or to "sanitize" the 5′ end of transcripts by removing out-of-frame AUG codons (Davis 1996). Finally, we note that several mechanisms of translational regulation, including microRNA- and protein-mediated events, act through interactions with the mRNA cap (de Moor et al. 2005; Kiriakidou et al. 2007). By providing a TMG cap (instead of MMG), SL *trans*-splicing may be affecting the receptiveness of mRNAs to different modes of translational regulation.

One of the most puzzling aspects of *trans*-splicing is its phylogenetic distribution (Nilsen 2001; Hastings 2005; Roy and Irimia 2009). SL *trans*-splicing has been described in groups as diverse as euglenozoans (euglenoids and trypanosomes), dinoflagellates, hydrozoan cnidarians, nematodes, flatworms, bdelloid rotifers, chaetognaths, and urochordates (Van der Ploeg 1986; Krause and Hirsh 1987; Rajkovic et al. 1990; Tessier et al. 1991; Stover and Steele 2001; Vandenberghe et al. 2001; Ganot et al. 2004; Blumenthal 2005; Pouchkina-Stantcheva and Tunnacliffe 2005; Satou et al. 2006; Zhang et al. 2007; Marletaz et al. 2008). But it has not yet been detected in other groups, including fungi, plants, vertebrates, and arthropods, which comprise some of the most intensively studied organisms. This distribution suggests two possibilities: either *trans*-splicing is an ancient mechanism that was independently lost in multiple lineages or it evolved repeatedly within the eukaryotes (Hastings 2005). Currently, there is no reliable way to decide between these alternatives. Most of the mechanisms and components used in SL *trans*-splicing are identical to those of *cis*-splicing, and the few that are unique to *trans*-splicing (SL RNA and accessory proteins) do not appear to be conserved between taxa (Nilsen 2001; Denker et al. 2002; Hastings 2005). Several authors have suggested that as new members are added to the list of organisms that possess *trans*-splicing, the hypothesis of ancient origins will become increasingly parsimonious (Nilsen 2001; Vandenberghe et al. 2001; Hastings 2005; Pouchkina-Stantcheva and Tunnacliffe 2005).

Here, we present compelling evidence for SL *trans*-splicing in amphipod crustaceans, representing the first evidence of this process in arthropods. A survey of expressed sequence tag (EST) data sets from a wide range of metazoans suggests that it also occurs in copepod crustaceans, ctenophores, and hexactinellid sponges. However, we find no evidence for SL *trans*-splicing in a number of well-studied arthropods, including close relatives of the amphipods, nor in the closest relatives of the hexactinellids—the calcarean sponges and demosponges. Furthermore, we find no evidence for SL *trans*-splicing in representatives of 14 other, previously unstudied, animal phyla. The incidence of SL *trans*-splicing in the metazoan phylogeny appears as fragmented as ever. We propose that SL *trans*-splicing probably evolved several times; we discuss a possible mechanism for the evolution of *trans*-splicing that might explain this puzzling evolutionary plasticity.

Materials and Methods

Analysis of Parhyale *trans*-Spliced ESTs

We first identified SL sequences as common leader (5′) sequences present in unrelated cDNA sequences from *Parhyale*, including GenBank sequences (accession numbers DQ827719, DQ827720, DQ827721, and FN568490) and other cDNA sequences identified in our laboratory. In subsequent analyses, we used the "trimmed and filtered" EST data set for *Parhyale hawaiensis* produced by the US Department of Energy Joint Genome Institute (JGI; v1.0, released May 2008), which includes 47,732 sequence reads; approximately, half of these are reads from the 5′ end of cDNAs.

To identify transcripts with leader sequences at their 5′ end, we searched the first (most 5′) 45 nucleotides of each EST for matches to each variant of the leader sequence. In all, 1673 matches were found with a ten-nucleotide query sequence (fig. 1*A*); this represents a frequency of 7% among 5′ EST reads. To estimate the real incidence of *trans*-splicing, we obtained a rough estimate of the completeness of 5′ ends in the JGI data set by comparing the ESTs with full-length cDNAs found in the GenBank database (accession numbers DQ917572, DQ917573, EU289288, EU289291, and EU289289). We found that out of 11 corresponding ESTs, none had a complete 5′ end, but seven were long enough to allow detection of a potential leader sequence using a ten-nucleotide query sequence. Based on these figures, we estimate that at least 10% of *Parhyale* transcripts carry a leader sequence.

To test whether individual mRNAs can combine with different SL variants, we focused on 122 distinct *trans*-spliced exons (minimum 20 nucleotides with 100% identity) that are represented by two or more *trans*-spliced ESTs in the JGI database (carrying a minimum of 13 nucleotides of each leader sequence, to allow discrimination between different SL variants). The majority of these mRNAs were associated with more than one SL variant (example shown in fig. 1*B*). Rare SL variants that might represent sequencing errors were not included in this analysis. The complete data are available on request.

To determine the splice acceptor motif, we searched the JGI data set for ESTs representing unspliced transcripts among mRNAs that are also found associated with an SL. We examined 104 distinct mRNAs for which there are ESTs both with and without a leader sequence in the JGI data set. In 60 out of 104 cases, leaderless ESTs had a YAG (CAG or TAG) sequence just 5′ of the splice junction; this sequence was often preceded by a T-rich stretch (figs. 1*C* and *D*). The putative splice acceptor motif shown in figure 1*D* was drawn from these 60 sequences using Pictogram (http://genes.mit.edu/pictogram.html). The remaining cases may represent *cis*-spliced sequences, derived from rare instances where *cis*-splicing acceptors have been used in spurious *trans*-splicing events (a significant proportion of these mRNAs have a start codon and long open-reading frame starting before the splice junction, and the putative *cis*-spliced products are more frequent that the *trans*-spliced ones).

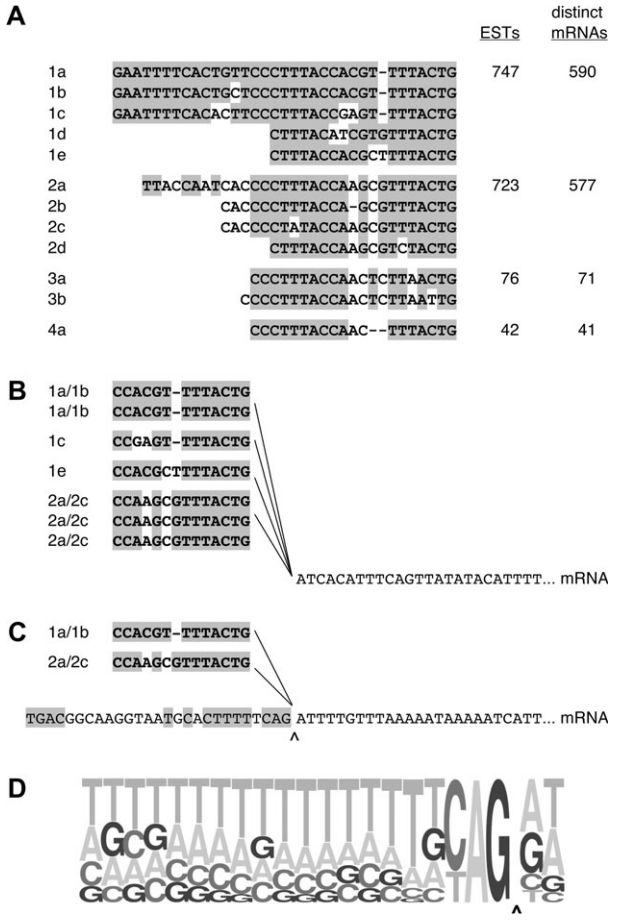

Fig. 1. SL *trans*-splicing in the amphipod crustacean *Parhyale hawaiensis*. (A) Major *Parhyale* SL variants, categorized by sequence similarity (gray highlights similarity), indicating the number of ESTs and the number of distinct mRNAs that carry each type of SL in the JGI data set. Eighty-five additional ESTs carry closely related sequences that represent rare SL variants or SLs with sequencing errors. (B) Example of ESTs representing an mRNA sequence that has been spliced to at least four different SL variants. (C) Example of ESTs representing an mRNA sequence in *trans*-spliced or unspliced forms. The unspliced form carries a putative splice branch point followed by a splice acceptor motif (highlighted in gray) just upstream of the splice site (arrowhead). (D) Consensus *trans*-splicing acceptor motif derived from 60 putative unspliced ESTs (see Materials and Methods); the height of letters represents the frequency of nucleotides at each position relative to the splice site (arrowhead). This motif is indistinguishable from the canonical *cis*-splicing motif.

Phylogenetic Survey of EST Data Sets

We downloaded raw ESTs of diverse organisms from the National Center for Biotechnology Information database or used curated ESTs from which contaminating vector sequence had been removed (http://www.estinformatics.org). To reduce the possibility of analyzing multiple copies of the same sequence, we used CAP3 (Huang and Madan 1999) to assemble ESTs from each species into contigs. Searches for common leader sequences were carried out on data sets containing 100 nucleotides from the 5′ end and from the reverse complemented 3′ end of all contigs and singleton ESTs, for each species.

We developed a Perl programme called Quickmatch to search for conserved ends of ESTs (up to 50 bp long) followed by divergent sequence. Quickmatch compared each 100-nucleotide sequence against all others. An initial 6-bp match was searched for and then extended if possible, allowing a degree of mismatch (one base in six was allowed to be a substitution or deletion/insertion, both in the initial seed hexamer and in the extension). Matches shorter than 12 bp or longer than 50 bp were rejected. For increased speed, once a pairwise match had been made, the matching sequence was not used as a query or interrogated again by later query sequences. All matches were then grouped and aligned by ClustalW (Larkin et al. 2007). The aligned putative sets of SL sequences were written to an html file colored according to nucleotide for easy identification of conserved regions of nucleotides.

The aligned sequences were scanned by eye to identify possible SLs among a much larger proportion of false positives (such as weak matches and homopolymer tracts). SLs were identified as stretches of conserved 5′ sequence followed by completely divergent sequence downstream. Recognizing the possibility that such sequences could represent primers used for amplifying the ESTs or vector multiple cloning sites, we analyzed potential SLs to rule out such false positives. First, we used Webcutter (http://rna.lundberg.gu.se/cutter2/) to examine possible SLs for restriction sites with the expectation that artificial sequences might have an excess of recognition sites for commonly used enzymes. Second, we used a Blast search against the VecScreen database (http://www.ncbi.nlm.nih.gov/VecScreen/) to look for hits against curated vector/primer sequences. Third, we used BlastN against metazoan EST data sets (excluding human, mouse, and the species under investigation) with the expectation that artificial sequences might be represented at the termini of ESTs of other species. Finally, we required that the putative SL be found uniquely at the 5′ ends of transcripts, as assessed by BlastX against the nonredundant protein database to indicate the orientation of the coding sequence.

Cloning of SL Precursor Genes and snRNA

SL repeats were amplified by polymerase chain reaction (PCR) on genomic DNA, using outward-facing primers designed for *Parhyale* SL1a (SLRF: 5′-CCTTTACCACGTTTTACTG-3′, SLRR: 5′-AAGGGAACAGTGAAAATTC-3′), *Mnemiopsis* SL1 (MnSL1F: 5′-CAACTACTATTAATTAAATAATTTGA-3′, MnSL1R: 5′-TTAATAGTAGTTGTTGAAAGTAT-3′), *Mnemiopsis* SL2 (MnSL2F: 5′-CTACAAATTAAATACATTTATTGAG-3′, MnSL2R: 5′-TTAATTTGTAGTGTTGAAATAGTT-3′), *Adineta* SL (AdSLF: 5′-TGCGATGACGAAAACGTGCGG-3′, AdSLR: 5′-CCTCTTGGTAAGTTGTAATAAGCC-3′), *Spadella* SL2.0 (ScSL20F: 5′-GAGTAGTTTCAATTTGTTTAAA-3′, ScSL20R: 5′-AACTACTCAATTATAAGCTTCC-3′), and *Calanus* SL (CfSL1F: 5′-CTTGAGTATAACACTTTAAAAGA-3′, CfSL1R: 5′-CAATATGAGTTCGTCACATCGAA-3′, CfSL3F: 5′-GCTTGTCTAAACACTTTAAAAGA-3′, CfSL3R: 5′-TTAGACAAGCAGTATAGCTTGG-3′). PCR was carried out on genomic DNA extracted from *P. hawaiensis*, *Mnemiopsis*

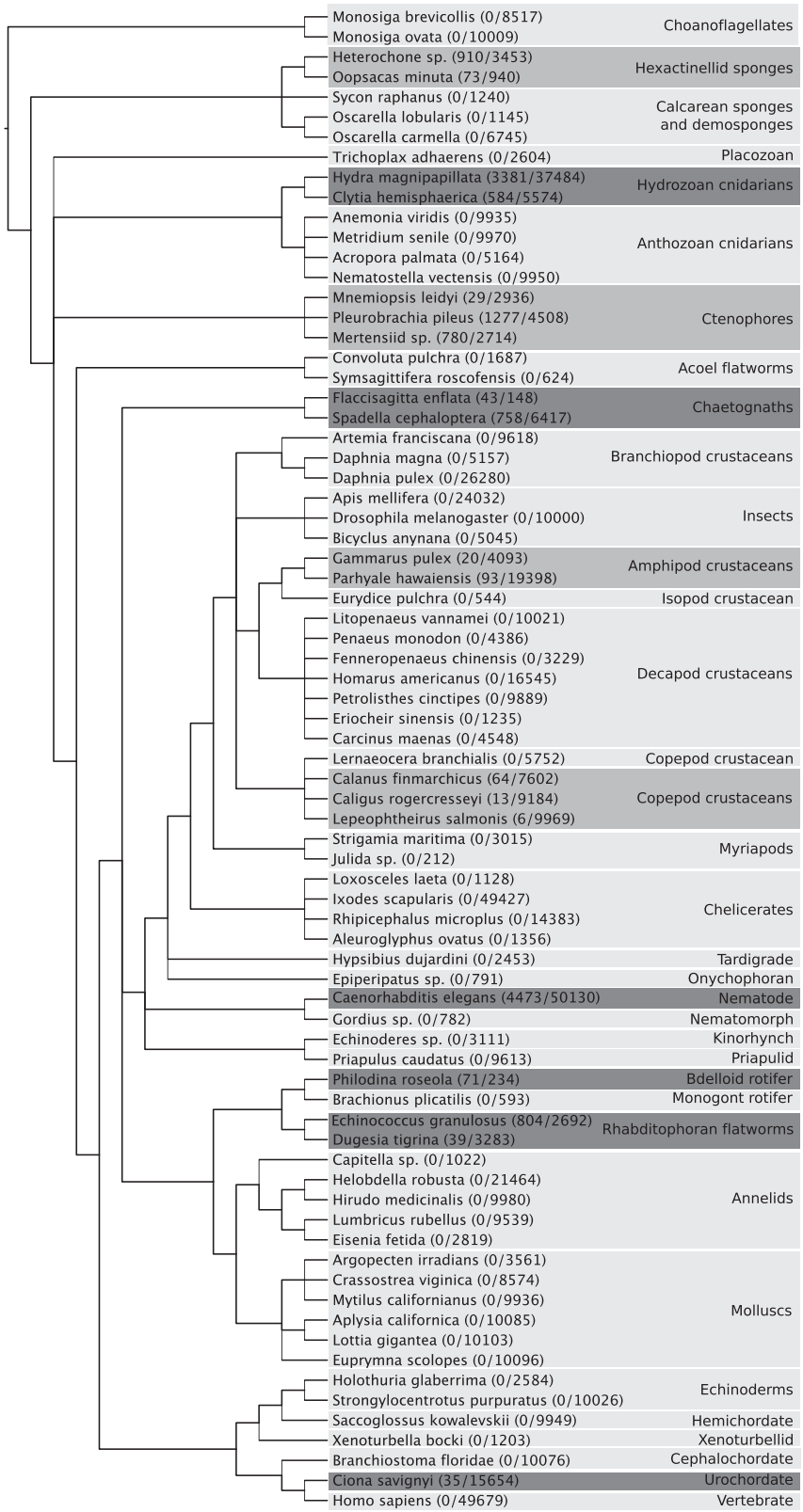

Fig. 2. Phylogenetic survey of SL *trans*-splicing in Metazoa, based on analysis of 75 EST data sets. The species where we have found SLs are marked by shading; species where SL *trans*-splicing was previously known are marked by darker shading. The incidence of SL *trans*-splicing in these EST data sets is shown for each species (number of EST contigs with SL sequences over total number of EST contigs analyzed). The incidence of *trans*-splicing in these data sets does not directly reflect the real frequency of *trans*-splicing in each species, as it also depends on the degree of completeness of 5′ ends in the corresponding cDNA samples and on EST sequencing strategy. These figures represent a minimum estimate for the frequency of SL *trans*-splicing in each species. Tree topology is based on consensus of recent molecular phylogenies (Telford 2006); branch lengths are arbitrary.

leidyi (gift from Kevin Pang and Mark Martindale), *Adineta ricciae* (gift from Chiara Boschetti and Alan Tunacliffe), *Spadella cephaloptera* (gift from Roxane Barthelemy), and *Calanus finmarchicus/helgolandicus* (gift from Jasmin Renz). Genomic repeats of SL precursor genes were recovered from *Parhyale*, *Mnemiopsis*, *Adineta*, and *Spadella* (accession numbers FN434129–FN434136).

Linkage between SL precursor genes and 5S ribosomal RNA (rRNA) in *Parhyale* was tested by PCR using pairwise combinations between the *Parhyale* SL primers (above) and outward-facing primers targeted to highly conserved regions of 5S rRNA (5SLeft: 5′-TAACTTCGCTGATCGGACGAGA-3′, 5SRight: 5′-GACCGCCTGGGAACACCAGATG-3′). The 0.6- and 0.8-kb repeats of 5S rRNA, and a 1.4-kb double repeat, were amplified using the 5S primer pair alone (accession number FN434137).

The *Parhyale* SL precursor RNA was amplified from total RNA by Rapid Amplification of cDNA Ends (RACE). Nucleic acids were isolated from *Parhyale* embryos, DNAase treated and polyadenylated using a poly-A tailing kit (Ambion). The SL RNA was then amplified using a primer targeting the known SL snRNA sequence (5′-CCCTTTACCACGTTTTACTGGTATCGTA-3′) and the 3′ primer from the SMART RACE kit (Clontech). The identity of this product was confirmed by sequencing (sequence accession number FN568489).

The putative secondary structure of SL precursor RNAs was determined using RNAfold and RNAalifold (Gruber et al. 2008). The *U1* snRNA gene in *Parhyale* was identified using Blast (Altschul et al. 1997).

Results

SL *trans*-Splicing in an Arthropod

While working with cDNA sequences from the amphipod crustacean *P. hawaiensis*, we noticed that several unrelated cDNAs carried a common sequence at their 5′ end. These included cDNAs from the Hox gene *Ultrabithorax (Ubx)*, of transcription factors *twist* and *mef2*, and of two exon-trapped genes cloned in our laboratory (Douris V and Averof M, unpublished data). To explore whether this short leader sequence might be present on a larger number of transcripts, we searched a database of the JGI that contains approximately 24,000 reads from the 5′ end of *Parhyale* cDNAs. We found more than 1,500 ESTs that contain variants of the same leader sequence at their 5′ end (fig. 1*A*). Most ESTs with a leader sequence correspond to distinct mRNAs (fig. 1*A*), suggesting that a major part of the *Parhyale* transcriptome—significant in both number and diversity of transcripts—shares this feature.

Although the leader sequences in these ESTs have a precisely defined 3′ end, their 5′ end is truncated to varying extents, indicating that most ESTs in this data set derive from cDNAs with incomplete 5′ ends. Due to these truncations, direct inspection of the EST data set is likely to underestimate the number of transcripts carrying a leader sequence. Taking into account the incompleteness of 5′ ends in the EST data set (see Materials and Methods), we estimate that at least 10% of *Parhyale* transcripts carry a leader sequence.

The existence of several variants of the leader sequence, with a few nucleotide differences (fig. 1*A*), allows us to test the hypothesis that leader sequences are added to mRNAs by SL *trans*-splicing: If a leader and mRNA are transcribed from the same locus, we expect that each gene will be associated with a single variant of the leader sequence; in contrast, if the leader is acquired by *trans*-splicing, we expect that each mRNA could combine with different variants from the pool of SL RNAs. In the JGI EST data set, we find clear evidence for the latter: Pairs of ESTs corresponding to the same mRNA are often associated with different variants of the leader sequence (49 of 82 pairs), and we found 15 examples of mRNAs that were associated with three or four different variants (e.g., fig. 1*B*). These findings support the idea that the leader sequences are acquired by SL *trans*-splicing.

Three additional lines of evidence confirm that these leader sequences are acquired by *trans*-splicing. First, fully sequenced bacterial artificial chromosome clones from the genomic regions of the *Parhyale twist* and *Ubx* genes show that the leader sequence is not contained within 79 or 132 kb, respectively, upstream of the *twist* and *Ubx* transcription units (Serano J, Hannibal R, and Patel NH, personal communication). Second, we have amplified SL precursor RNAs that carry the leader sequence and bear structural similarities to other characterized SL snRNAs (see below). Third, using reverse transcriptase–PCR, we were able to detect splicing of the leader sequence to a splice acceptor site in the 5′ untranslated region of an exogenous construct introduced into the *Parhyale* genome by transgenesis (*PhHS–DsRed* construct; Pavlopoulos et al. 2009).

To determine the splice acceptor sequences that participate in *trans*-splicing events, we identified ESTs corresponding to putative unspliced transcripts (e.g., fig. 1*C*). The majority of these were found to contain a YAG (CAG or TAG) sequence just 5′ of the putative splice junction, often preceded by a T-rich stretch (fig. 1*D*). These features are reminiscent of the canonical splice acceptor motifs that have been defined for both *cis*- and *trans*-splicing (Conrad et al. 1993).

Phylogenetic Survey for SL *trans*-Splicing Reveals a Fragmented Distribution in Metazoans

Mapping the phylogenetic distribution of SL *trans*-splicing at higher resolution is important for understanding its evolution. To achieve this, we have analyzed 75 EST sequence data sets from diverse metazoan taxa (fig. 2), searching for common sequences (between 12 and 50 nucleotides in length) at 5′ ends of multiple unrelated transcripts. All such sequences were evaluated carefully to rule out the possibility that they represent vector sequences or primers used for amplifying the ESTs (see Materials and Methods).

We first examined ESTs from the amphipod *Gammarus pulex*, the closest relative of *Parhyale* represented in sequence databases; we found a convincing SL that is

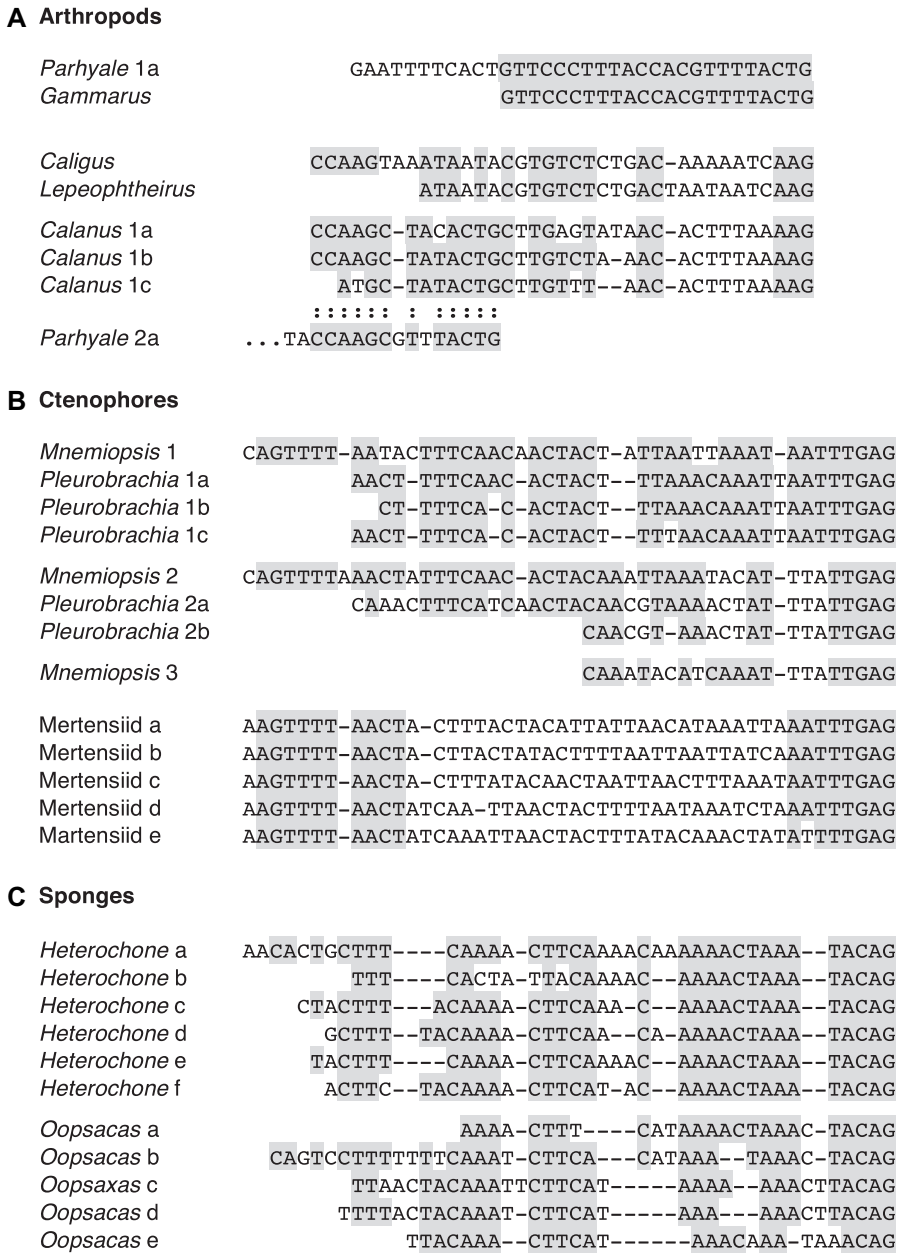

Fig. 3. Putative SL sequences discovered in arthropods, ctenophores, and sponges. Sequence similarities among SL variants within each group are highlighted in gray. (A) The SL found in the amphipod crustacean *Gammarus* is identical to *Parhyale* SL1a. Among copepod crustaceans, the closely related parasitic copepods *Caligus* and *Lepeophtheirus* have very similar SL sequences, which are more distantly related to those of the free-living copepod *Calanus*. The 5′ end of copepod SLs is similar to the 3′ end of amphipod SLs, particularly to *Parhyale* SL2; a detailed list of *Parhyale* SLs is given in figure 1A. (B) At least two distinct types of SLs (SL1 and SL2) are shared between the ctenophores *Mnemiopsis* and *Pleurobrachia*; a third type (SL3), with intermediate characteristics, was found in *Mnemiopsis*. The mertensiid ctenophore has a large number of SL variants with conserved 5′ and 3′ regions and divergent sequences in the middle. (C) The Hexactinellid sponges have a large number of SL variants (>50 per species) that are related to each other. Only a few common variants are shown for mertensiid, *Heterochone*, and *Oopsacas*. No obvious sequence similarities were found in SL comparisons between phyla.

identical to one of the SL variants found in *Parhyale* (fig. 3*A*). To test whether this SL is restricted to amphipods or shared more widely among crustaceans, we examined EST data sets from an isopod and seven decapods—belonging to the same class of crustaceans as amphipods, the Malacostraca—and found no evidence of SLs. Then, we examined a wider range of arthropods, including copepod and branchiopod crustaceans, insects, chelicerates (araneids and acarids), a myriapod, an onychophoran, and a tardigrade. We found a credible instance of SLs in three species of copepods but no plausible SLs among the other arthropod groups that we examined. The sequence near the 5′ end of copepod SLs shows some similarity to the 3′ end of *Parhyale* SLs (fig. 3*A*). Although this similarity might suggest a shared ancestry of amphipod and copepod SLs, it is difficult to explain how this sequence would have migrated along the length of the SL sequence.

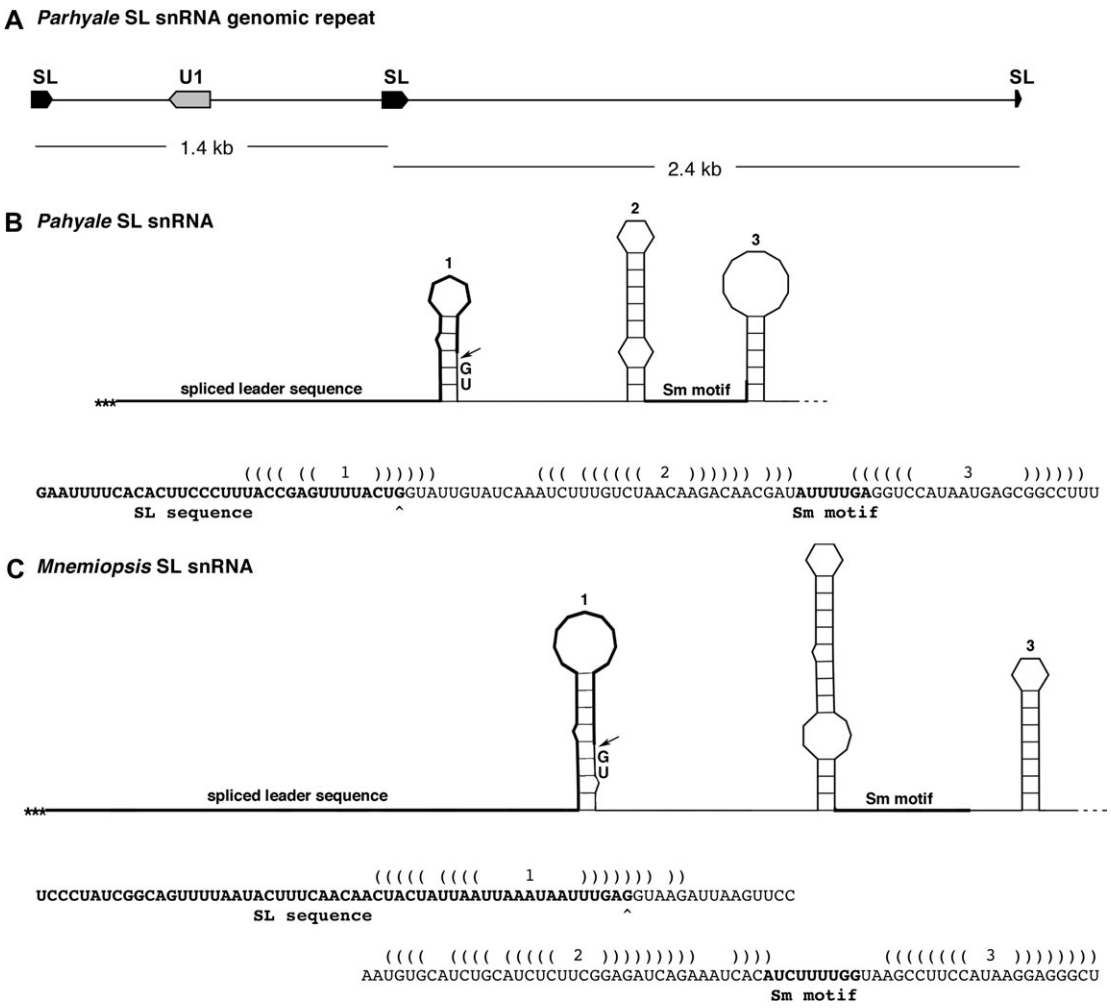

FIG. 4. SL precursor RNAs of amphipod crustaceans (*Parhyale*) and ctenophores (*Mnemiopsis*). (A) Genomic organization of SL snRNA gene repeats in *Parhyale*. Consecutive 1.4- and 2.4-kb tandem repeats were identified. The 1.4-kb repeats contain a U1 snRNA gene in opposite orientation to the SL genes. (B and C) Predicted secondary structure of SL precursor RNAs of *Parhyale* and *Mnemiopsis*, indicating the position of the SL sequence, the splice donor site (arrow, with characteristic UG dinucleotide immediately downstream), and the putative Sm motif, in relation to the predicted stem–loop structures 1, 2, and 3. In the sequence below each diagram, we indicate the nucleotides participating in each stem–loop (brackets above sequence), the splice site (arrowhead), and the SL sequence and Sm motif (in bold). The extent of the 5′ end of the *Mnemiopsisis* SL RNA is deduced from sequence conservation among divergent genes for SL1 and SL2; this extended 5′ sequence contains a second putative Sm motif.

Extending our phylogenetic reach to other Ecdysozoa, we examined the nematomorphs (likely sister groups of the nematodes, in which SLs are prevalent), a kinorhynch, and a priapulid worm; we were unable to find likely SLs in any of these species. Among other metazoan phyla, we replicated the finding of SLs in rhabditophoran flatworms and chaetognaths but not in acoelomorphs (Marletaz et al. 2008). Our software also found the known SLs in the bdelloid rotifer *Philodina* (Pouchkina-Stantcheva and Tunnacliffe 2005) but found none in the monogont rotifer *Brachionus*. Finally, we found strong evidence for previously unreported SLs in the diploblastic ctenophores (sea gooseberries) and hexactinellid sponges. We found no evidence for SLs in species surveyed from the annelids, molluscs, kinorhynchs, priapulids nematomorphs, tardigrades, onychophorans, xenoturbellids, hemichordates, echinoderms, acoelomorphs, placozoans, or choanoflagellates (see fig. 2).

Our ability to detect *trans*-splicing in data sets with a low incidence of SLs (e.g., less than 1% in copepods) and in the species where SL *trans*-splicing was previously documented gives us confidence that our method is sensitive. Nevertheless, it remains impossible to disprove the existence of SL *trans*-splicing in any given taxon; some EST data sets are small, some species have very low levels of SL *trans*-splicing, and the vast majority of living species have not been sampled.

Identification of SL Precursor Genes

To show that SL sequences derive from independent genes, distinct from the diverse mRNAs they precede, we set out to clone the genes coding for SL precursor RNAs in several species. Previous studies have shown that SL precursor genes are usually arranged in tandem repeats, often located within the repeats of 5S rRNA or of other snRNA genes

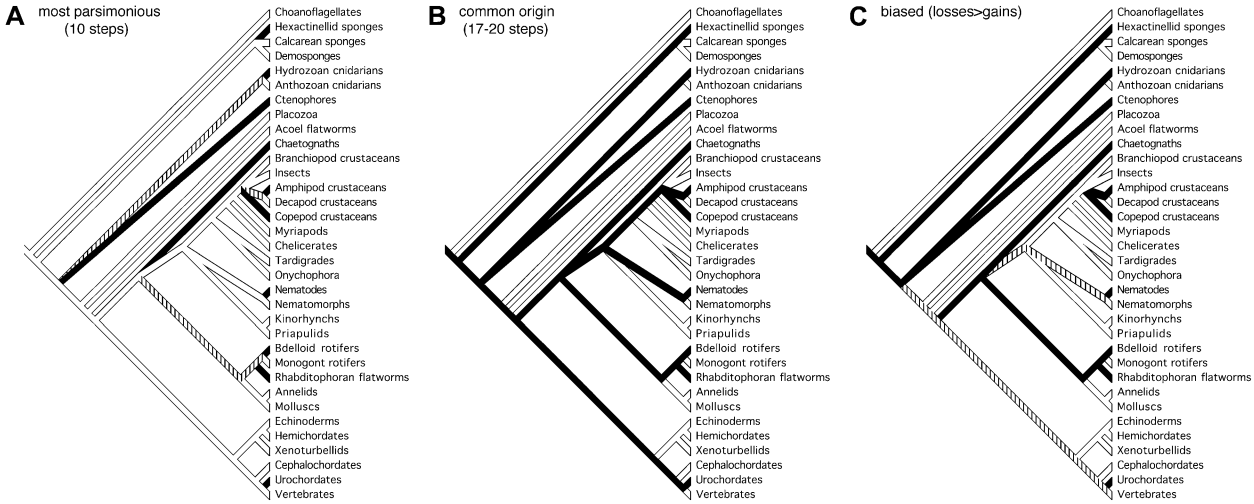

Fig. 5. Alternative scenarios for the evolution of SL *trans*-splicing in Metazoa. (A) The most parsimonious interpretation of our data suggests that SL *trans*-splicing evolved repeatedly within the animal kingdom from a common ancestor that did not have this capacity; six to ten independent gains of SL *trans*-splicing are predicted with equal parsimony (total number of gains and losses is ten). (B) An evolutionary scenario assuming a single origin of *trans*-splicing would require 17–20 independent losses of *trans*-splicing in the animals and choanoflagellates (precise number of losses depends on the resolution of phylogenetic uncertainties, indicated as polytomies). (C) If we assume a bias in the polarity of changes, where losses of *trans*-splicing are twice as likely as gains, the most parsimonious reconstruction still predicts the two to five independent origins of SL *trans*-splicing within the Metazoa. Solid black branches indicate lineages where SL *trans*-splicing is inferred to be present, white branches indicate lineages where it is inferred to be absent, and hatched branches indicate lineages where the presence or absence of SL *trans*-splicing are equally parsimonious.

(Drouin and de Sa 1995; Ebel et al. 1999; Stover and Steele 2001; Ganot et al. 2004). Assuming that this organization may be found in other species, we designed pairs of primers to amplify across the putative SL repeats, targeting the amphipod, copepod, and ctenophore SLs reported in this study, as well as chaetognath and bdelloid rotifer SLs (previously reported but not cloned from the respective genomes; Pouchkina-Stantcheva and Tunnacliffe 2005; Marletaz et al. 2008).

Using this approach, we amplified 1.4- and 2.4-kb fragments corresponding to single SL precursor repeats from the genome of the crustacean *P. hawaiensis*, which combine to give a double-repeat unit of 3.8 kb (fig. 4*A*). The gene cluster was found to contain a U1 snRNA gene but no sequences corresponding to 5S rRNA. To test whether there may be loci in which SL precursors are linked to 5S rRNA genes in *Parhyale*, we carried out PCR with combinations of primers targeting SL RNA and 5S rRNA sequences. We cloned 5S rRNA repeats but failed to amplify any fragment containing both 5S rRNA and SL precursor sequences. Thus, in *Parhyale*, SL precursor genes are associated with U1 snRNA genes, but there is no evidence for linkage with 5S rRNA genes. We also used 3′ RACE on *Parhyale* RNA to amplify the corresponding SL precursor RNA.

Using the same strategy to amplify across the adjacent SL repeats, we were able to clone 0.9- and 0.4-kb tandem repeat units of SL1 and SL2 from the genome of the ctenophore *M. leidyi*, 0.5- and 0.6-kb tandem repeats of SL2 from the chaetognath *S. cephaloptera*, and a 1-kb inverted repeat of an SL gene from the bdelloid rotifer *A. ricciae*. In these cases, we did not find any 5S rRNA, snRNA, or other genes associated with these gene clusters. We were not able to amplify SL precursor genes from the copepod *C. finmarchicus*.

The sequences of the SL precursors that we recovered contain features that are similar to those described in SL precursor RNAs of other species (Bruzik et al. 1988; Davis 1996). In all cases, the SL sequences have the characteristic GU dinucleotide just 3′ to the splice junction. Conceptual folding of the *Parhyale* and *Mnemiopsis* SL RNAs suggests that these may adopt a secondary structure consisting of three consecutive stem–loops, with the splice donor site located in a base-paired region at the 3′ end of stem–loop 1 and a putative Sm motif located between stem–loops 2 and 3 (figs. 4*B* and *C*). Besides these features, which are likely to reflect basic requirements for snRNP assembly and *trans*-splicing (Hannon et al. 1991; Will and Luhrmann 2001), the SL precursor sequences that we have identified show no obvious sequence similarity to previously identified SL precursors from other species.

Discussion

The discovery of *trans*-splicing in different phyla has, until now, been based on serendipity; researchers working with cDNA sequences in a species of interest have accidentally observed the presence of common 5′ sequences in unrelated transcripts and then determined that these leader sequences and the corresponding mRNAs are transcribed from distinct genomic loci. This haphazard approach has meant that the identification of SL *trans*-splicing in different phyla has not been systematic and negative results have not been published. The increasing availability of EST sequences from a wide range of animals now provides the opportunity to survey the phylogenetic distribution of SL *trans*-splicing in metazoans.

Prompted by the accidental discovery of SL *trans*-splicing in an amphipod crustacean (*Parhyale*), we have carried out a survey of 75 EST data sets from 23 metazoan phyla and their sister group, the choanoflagellates. Our survey provides the first evidence that SL *trans*-splicing also occurs in three species of copepod crustaceans, three ctenophores, two hexactinellid sponges, and an additional species of amphipod. No evidence for *trans*-splicing could be found in the closest relatives of the amphipods (other malacostracans, including decapod and isopod crustaceans) or in more distantly related arthropod groups. Thus, in spite of the apparent similarity between amphipod and copepod SLs (fig. 3*A*), our survey suggests that the occurrence of *trans*-splicing in amphipods and copepods is not part of a broader distribution of the phenomenon within the arthropods. Similarly, at the base of the metazoan tree, the discovery of SL *trans*-splicing in hydrozoan cnidarians (Stover and Steele 2001), hexactinellid sponges, and ctenophores might suggest that this mechanism was present in the ancestors of metazoans (Philippe et al. 2009), yet no evidence for *trans*-splicing is found in anthozoan cnidarians, demosponges, calcarean sponges, or a placozoan. The putative SL sequences that our survey has uncovered show no obvious sequence similarities across phyla.

The results of our phylogenetic survey are summarized in figure 2. Overall, the survey shows that although a denser sampling of the metazoan phylogeny has revealed new instances of *trans*-splicing, this is matched by an increase in the number of taxa that lack *trans*-splicing. The increased resolution does not result in a more cohesive phylogenetic distribution of SL *trans*-splicing. The most parsimonious reconstruction of SL evolution according to our data would require ten independent instances of gain or loss of *trans*-splicing—with at least six independent gains in the hexactinellids, ctenophores/cnidarians, crustaceans, nematodes, chaetognaths/rotifers/rhabditophorans, and urochordates—from a common ancestor that did not have the capacity to *trans*-splice (fig. 5*A*). The alternative scenario, assuming a single origin of *trans*-splicing, would require 17–20 independent losses of *trans*-splicing (fig. 5*B*). Even if we assume that losses of *trans*-splicing are twice as likely as gains, our study predicts two to five independent origins of SL *trans*-splicing in Metazoa (fig. 5*C*).

The hypothesis that *trans*-splicing arose independently in multiple groups requires that such a mechanism could evolve relatively easily. If new components of splicing machinery and complex splicing reactions had to evolve de novo, multiple origins would appear implausible. Current knowledge suggests that *trans*-splicing uses the same molecular machinery as *cis*-splicing, except that in *trans*-splicing U1 snRNP is missing and is replaced by an snRNP formed by the SL precursor RNA (Bruzik et al. 1988; Bruzik and Steitz 1990; Hannon et al. 1991). Strikingly, it is possible to transfer the capacity for *trans*-splicing to organisms that do not have that ability simply by providing an SL precursor RNA (Bruzik and Maniatis 1992), which suggests that the evolution of an SL precursor RNA may be the only requirement for acquiring the capacity to carry out *trans*-splicing reactions. The only evidence for factors that are required specifically for *trans*-splicing comes from *Caenorhabditis* (Denker et al. 2002), but these factors appear not to be widely conserved.

How might an SL precursor evolve? SL precursors are small RNA molecules with a splice donor site, an Sm-binding motif, and a secondary structure that will allow the assembly of an snRNP. The latter two features are shared with the U1, U2, U4, and U5 spliceosomal snRNAs, and the mutually exclusive relationship of the U1 and SL snRNPs could be taken to suggest that the SL precursor is homologous to or derived from U1. In support of this idea, experiments have shown that it is possible to convert a U1 snRNA into an SL precursor simply by adding a splice donor and changing a few nucleotides next to the Sm motif (Hannon et al. 1992). These experiments suggest that it may just take a few steps to evolve an SL precursor from existing spliceosomal snRNAs or from other small RNAs. The location of SL precursors in a gene cluster that also contains the U1 snRNA in *Parhyale* (fig. 4*A*) provides some support for this hypothesis in the case of amphipod crustaceans. Taking these considerations into account, the repeated evolution of *trans*-splicing within the animal kingdom appears to be an increasingly plausible scenario.

Acknowledgments

We thank Nipam Patel and the Joint Genome Institute for making the *Parhyale* EST data publicly available; the *Parhyale* cDNA libraries and EST sequences were generated by Julia Serano, Jay Rehm, Matthias Gerberding, Bill Browne, Alivia Price, Ehab Abouheif, Matt Giorgianni, Erika Lindquist, Peter Brokstein, and Nipam Patel. We also thank Tassos Pavlopoulos, Alexandros Kiupakis, Nikos Konstantinides, and Haris Kontarakis for generating and characterizing gene traps that lead to the discovery of *trans*-splicing in *Parhyale*; Roxane Barthelemy, Chiara Boschetti and Alan Tunnacliffe, Kevin Pang and Mark Martindale, Andreas Hejnol, and Jasmin Renz for sending DNA or preserved specimens from diverse species; Juan Valcarcel and Maura Strigini for discussion; and the Marie Curie Research Training Network "ZOONET" (European Union, Framework Programme 6) for fostering interaction between our laboratories.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

Blumenthal T. 2005. Trans-splicing and operons. *WormBook*. doi:10.1895/wormbook.1.5.1. 1–9.

Bonen L. 1993. Trans-splicing of pre-mRNA in plants, animals, and protists. *FASEB J.* 7:40–46.

Bruzik JP, Maniatis T. 1992. Spliced leader RNAs from lower eukaryotes are trans-spliced in mammalian cells. *Nature* 360:692–695.

Bruzik JP, Steitz JA. 1990. Spliced leader RNA sequences can substitute for the essential 5′ end of U1 RNA during splicing in a mammalian in vitro system. *Cell* 62:889–899.

Bruzik JP, Van Doren K, Hirsh D, Steitz JA. 1988. Trans splicing involves a novel form of small nuclear ribonucleoprotein particles. *Nature* 335:559–562.

Cheng G, Cohen L, Ndegwa D, Davis RE. 2006. The flatworm spliced leader 3′-terminal AUG as a translation initiator methionine. *J Biol Chem*. 281:733–743.

Conrad R, Liou RF, Blumenthal T. 1993. Conversion of a trans-spliced *C. elegans* gene into a conventional gene by introduction of a splice donor site. *EMBO J*. 12:1249–1255.

Davis RE. 1996. Spliced leader RNA trans-splicing in Metazoa. *Parasitol Today* 12:33–40.

de Moor CH, Meijer H, Lissenden S. 2005. Mechanisms of translational control by the 3' UTR in development and differentiation. *Semin Cell Dev Biol*. 16:49–58.

Denker JA, Maroney PA, Yu YT, Kanost RA, Nilsen TW. 1996. Multiple requirements for nematode spliced leader RNP function in trans-splicing. *RNA*. 2:746–755.

Denker JA, Zuckerman DM, Maroney PA, Nilsen TW. 2002. New components of the spliced leader RNP required for nematode trans-splicing. *Nature* 417:667–670.

Drouin G, de Sa MM. 1995. The concerted evolution of 5S ribosomal genes linked to the repeat units of other multigene families. *Mol Biol Evol*. 12:481–493.

Ebel C, Frantz C, Paulus F, Imbault P. 1999. Trans-splicing and cis-splicing in the colourless Euglenoid, *Entosiphon sulcatum*. *Curr Genet*. 35:542–550.

Ganot P, Kallesoe T, Reinhardt R, Chourrout D, Thompson EM. 2004. Spliced-leader RNA trans splicing in a chordate, *Oikopleura dioica*, with a compact genome. *Mol Cell Biol*. 24:7795–7805.

Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL. 2008. The Vienna RNA websuite. *Nucleic Acids Res*. 36:W70–W74.

Hannon GJ, Maroney PA, Nilsen TW. 1991. U small nuclear ribonucleoprotein requirements for nematode cis- and trans-splicing in vitro. *J Biol Chem*. 266:22792–22795.

Hannon GJ, Maroney PA, Yu YT, Hannon GE, Nilsen TW. 1992. Interaction of U6 snRNA with a sequence required for function of the nematode SL RNA in trans-splicing. *Science* 258:1775–1780.

Hastings KE. 2005. SL trans-splicing: easy come or easy go? *Trends Genet*. 21:240–247.

Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. *Genome Res*. 9:868–877.

Kiriakidou M, Tan GS, Lamprinaki S, De Planell-Saguer M, Nelson PT, Mourelatos Z. 2007. An mRNA m7G cap binding-like motif within human Ago2 represses translation. *Cell* 129:1141–1151.

Krause M, Hirsh D. 1987. A trans-spliced leader sequence on actin mRNA in *C. elegans*. *Cell* 49:753–761.

Lall S, Friedman CC, Jankowska-Anyszka M, Stepinski J, Darzynkiewicz E, Davis RE. 2004. 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*. 279:45573–45585.

Larkin MA, Blackshields G, Brown NP, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.

Lee MG, Van der Ploeg LH. 1997. Transcription of protein-coding genes in trypanosomes by RNA polymerase I. *Annu Rev Microbiol*. 51:463–489.

Maniatis T, Tasic B. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418:236–243.

Marletaz F, Gilles A, Caubit X, Perez Y, Dossat C, Samain S, Gyapay G, Wincker P, Le Parco Y. 2008. Chaetognath transcriptome reveals ancestral and unique features among bilaterians. *Genome Biol*. 9:R94.

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

Maroney PA, Hannon GJ, Denker JA, Nilsen TW. 1990. The nematode spliced leader RNA participates in trans-splicing as an Sm snRNP. *EMBO J*. 9:3667–3673.

Nilsen TW. 2001. Evolutionary origin of SL-addition trans-splicing: still an enigma. *Trends Genet*. 17:678–680.

Pavlopoulos A, Kontarakis Z, Liubicich DM, Serano JM, Akam M, Patel NH, Averof M. 2009. Probing the evolution of appendage specialization by Hox gene misexpression in an emerging model crustacean. *Proc Natl Acad Sci U S A*. 106:13897–13902.

Philippe H, Derelle R, Lopez P, et al. 2009. Phylogenomics revives traditional views on deep animal relationships. *Curr Biol*. 19:706–712.

Pouchkina-Stantcheva NN, Tunnacliffe A. 2005. Spliced leader RNA-mediated trans-splicing in phylum Rotifera. *Mol Biol Evol*. 22:1482–1489.

Rajkovic A, Davis RE, Simonsen JN, Rottman FM. 1990. A spliced leader is present on a subset of mRNAs from the human parasite *Schistosoma mansoni*. *Proc Natl Acad Sci U S A*. 87:8879–8883.

Roy SW, Irimia M. 2009. Splicing in the eukaryotic ancestor: form, function and dysfunction. *Trends Ecol Evol*. 24:447–455.

Satou Y, Hamaguchi M, Takeuchi K, Hastings KE, Satoh N. 2006. Genomic overview of mRNA 5′-leader trans-splicing in the ascidian *Ciona intestinalis*. *Nucleic Acids Res*. 34:3378–3388.

Stover NA, Steele RE. 2001. Trans-spliced leader addition to mRNAs in a cnidarian. *Proc Natl Acad Sci U S A*. 98:5693–5698.

Telford MJ. 2006. Animal phylogeny. *Curr Biol*. 16:R981–R985.

Tessier LH, Keller M, Chan RL, Fournier R, Weil JH, Imbault P. 1991. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in *Euglena*. *EMBO J*. 10:2621–2625.

Ploeg Van der LH. 1986. Discontinuous transcription and splicing in trypanosomes. *Cell* 47:479–480.

Van Doren K, Hirsh D. 1990. mRNAs that mature through trans-splicing in *Caenorhabditis elegans* have a trimethylguanosine cap at their 5' termini. *Mol Cell Biol*. 10:1769–1772.

Vandenberghe AE, Meedel TH, Hastings KE. 2001. mRNA 5′-leader trans-splicing in the chordates. *Genes Dev*. 15:294–303.

Will CL, Luhrmann R. 2001. Spliceosomal UsnRNP biogenesis, structure and function. *Curr Opin Cell Biol*. 13:290–301.

Zeiner GM, Sturm NR, Campbell DA. 2003. The *Leishmania tarentolae* spliced leader contains determinants for association with polysomes. *J Biol Chem*. 278:38269–38275.

Zhang H, Hou Y, Miranda L, Campbell DA, Sturm NR, Gaasterland T, Lin S. 2007. Spliced leader RNA trans-splicing in dinoflagellates. *Proc Natl Acad Sci U S A*. 104:4618–4623.