Spatial and temporal preferences for *trans*-splicing in *Ciona intestinalis* revealed by EST-based gene expression analysis

Keywords: EST, regulation, ascidian, sequence analysis

Nicolas Siervo^{1,2}, Shuang Li^{1,3}, Yutaka Suzuki³, Riu Yamashita^{1,4}, and Kenta Nakai^{1,2,3,*}

1. Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.

2. Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST).

3. Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan.

4. Frontier Research Initiative, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.

*Correspondence should be addressed to:

Kenta Nakai, Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan. e-mail: knakai@ims.u-tokyo.ac.jp, TEL: +81 3 5449 5131, FAX: +81 3 5449 5133

Abbreviations: cDNA: DNA complementary to RNA; PCR: polymerase chain reaction; DDBJ: DNA data bank of Japan; EST: expressed sequence tag; GO: gene ontology; NCBI: national center for biotechnology information; SL: spliced-leader; SL(+): *trans*-spliced

5'-EST or gene; SL(-): non-*trans*-spliced 5'-EST or gene; SL(+/-): gene with both *trans*-spliced and non-*trans*-spliced 5'-ESTs; UCSC: university of California, Santa Cruz.

Abstract

Ciona intestinalis is a useful model organism to analyze chordate development and genetics. However, unlike vertebrates, it shares a unique mechanism called *trans*-splicing with lower eukaryotes. In the computational analysis of *trans*-splicing in *C. intestinalis* we report here, we discovered that although the amount of non-*trans*-spliced and *trans*-spliced genes is usually equivalent, the expression ratio between the two groups varies significantly with tissues and developmental stages. Among the seven tissues studied, the observed ratios ranged from 2.53 in “gonad” to 19.53 in “endostyle”, and during development they increased from 1.68 at the “egg” stage to 7.55 at the “juvenile” stage. We further hypothesize that this enrichment in *trans*-spliced mRNAs in early developmental stages might be related to the abundance of *trans*-spliced mRNAs in “gonad”. Our analysis indicates that in *C. intestinalis*, although there may not exist strong fundamental requirements for genes to be *trans*-spliced, the populations of non-*trans*-spliced and *trans*-spliced genes are likely to be spatially and temporally regulated differently.

1. Introduction

The intermolecular ligation of sequences from two different RNA molecules is defined as *trans*-splicing. In the case of mRNA spliced-leader *trans*-splicing, the splicing reaction results in the substitution of the 5'-end of a precursor mRNA by a short non-coding spliced-leader (SL) donated by a precursor SL RNA (Bonen, 1993; Vandenberghe et al., 2001). Because mRNA SL *trans*-splicing occurs at the 5'-end of mRNA, understanding it is necessary to locate accurate transcriptional start sites for gene regulation analysis.

Observation of mRNA SL *trans*-splicing has been reported in phyla broadly distributed throughout the evolutionary tree, including euglenozoans, nematodes, platyhelminthes (flatworms), cnidarians, rotifers and urochordates (Sutton and Boothroyd, 1986; Krause and Hirsh, 1987; Stover and Steele, 2001; Yuasa et al., 2002; Ganot et al., 2004; Pouchkina-Stantcheva and Tunnacliffe, 2005). Intensive studies of mRNA spliced-leader *trans*-splicing focusing on its mechanisms, functions and evolution (Krause and Hirsh, 1987; Laird, 1989; Nilsen, 1995; Blaxter and Liu, 1996; Hastings, 2005) have shown that although the basic principles remain unchanged, mRNA SL *trans*-splicing does not necessarily fulfill the same functions in all organisms. Several distinct roles have been reported, such as the resolution of polycistronic pre-mRNAs, which was observed to proceed via different mechanisms in kinetoplastids, nematodes, flatworms and urochordates (Davis and Hodgson, 1997; Blumenthal et al., 2002; Campbell et al., 2003; Satou et al., 2006); the enhancement of

translation by the SL, which occurs in kinetoplastids and nematodes (Maroney et al., 1995; Zeiner et al., 2003; Lall et al., 2004); the 5'-capping of protein-coding RNAs transcribed by RNA polymerase I in kinetoplastids (Lee and Van der Ploeg, 1997; Gunzl et al., 2003); or the sanitizing of 5' untranslated regions of pre-mRNA in metazoan (Davis, 1996).

In urochordates, mRNA SL *trans*-splicing was first identified in the ascidian *Ciona intestinalis* (Vandenberghe et al., 2001) and was then observed in *Halocynthia roretzi* and *Oikopleura dioica* (Yuasa et al., 2002; Ganot et al., 2004). The observation of mRNA SL *trans*-splicing in *C. intestinalis* is of particular interest because it shares basic features such as the chordate body plan and gene repertoires with vertebrates (Satoh and Jeffery, 1995; Corbo et al., 2001; Satou et al., 2001; Satoh, 2003; Satoh et al., 2003). In addition to its position in the evolutionary tree, a small genome size and a well-established cell lineage (Conklin, 1905; Nishida, 1987), as well as a transparent body and the availability of simple and efficient gene expression methods, contributed to the wide use of *C. intestinalis* in the study of chordate development (Passamanek and Di Gregorio, 2005).

The first observation of mRNA SL *trans*-splicing in *C. intestinalis* was made in 2001 by Vandenberghe et al., who reported the mRNA SL *trans*-splicing of at least seven genes by a 16-nucleotide SL (5'-ATTCTATTTGAATAAG-3') apparently derived from a 46-nucleotide precursor RNA (Vandenberghe et al., 2001). The later observations in *Halocynthia roretzi* and *Oikopleura dioica* showed that these two organisms have longer SLs (24- and

40-nucleotide SL respectively) (Yuasa et al., 2002; Ganot et al., 2004). Furthermore, while the SL of the two ascidians *C. intestinalis* and *H. roretzi* are very similar, with the exception of two short indels, that of the appendicularian *O. dioica* differs considerably (Vandenberghe et al., 2001; Yuasa et al., 2002; Ganot et al., 2004). In this organism, at least 25% of the mRNAs are *trans*-spliced, and this process is used to resolve polycistronic transcripts (Ganot et al., 2004). In *C. intestinalis*, a later study based on 2,078 oligo-capped ESTs representing 668 genes showed that the mRNAs transcribed from around 50% of the expressed genes undergo mRNA SL *trans*-splicing (Satou et al., 2006). Additionally, it was observed that, on average, *trans*-spliced mRNAs are 2 to 3 fold less abundant than non-*trans*-spliced mRNAs, and that ribosomal protein genes preferentially yielded non-*trans*-spliced mRNAs (Satou et al., 2006). Finally, unlike in Clade V nematodes where different SLs are normally used for mRNA SL *trans*-splicing and polycistronic pre-mRNA resolution (Blumenthal and Gleason, 2003), a unique SL is apparently involved in both functions in *O. dioica* and *C. intestinalis* (Ganot et al., 2004; Satou et al., 2006).

Interestingly, although it has been noted that the small fraction of the *Caenorhabditis* SL *trans*-spliced mRNAs that are derived from polycistronic operon transcripts largely encode ubiquitously expressed housekeeping proteins, mRNA SL *trans*-splicing does not appear to be specifically associated with particular tissues, developmental stages (Hastings, 2005). We report here the first observation of such associations, based on the clustering of 5'-ESTs with

all *C. intestinalis* ESTs included in UniGene build #19, a method frequently used as a measure for gene expression level since the early nineties and the analysis of the relationship between ESTs and gene expression in human (Adams et al., 1991; Wilcox et al., 1991).

2. Materials and methods

2.1 5'-EST sequences

We used 2,859 5'-EST sequences obtained by the oligo-capping method (Maruyama and Sugano, 1994), thus guaranteeing the correct sequencing of the mRNA 5'-end. 781 5'-ESTs were obtained from the whole body of an adult individual using a 454 DNA sequencer (small set) (Supplemental data 1, DDBJ accession numbers: XXXX), and 2,078 5'-ESTs were obtained from randomly picked cDNA clones generated from a mixture of egg, tailbud embryo, larval and young adult mRNAs (large set) (DDBJ accession numbers: BW648671-BW650748) (Satou et al., 2006).

For the small set, RNA was extracted using RNeasy (Qiagen) and 200 µg of the obtained total RNA was subjected to oligo-capping with some modifications from the original protocol (Maruyama and Sugano, 1994); BAP-TAP-treated total RNA was ligated with 1.2 µg of RNA oligo (5'-AGCAUCGAGUCGGCCUUGUUGGCCUUGCCAGCCCGCUCAGUUUC-3'). After DNase I (TaKaRa) treatment, RNAs ending with polyA were selected using oligo-dT powder (Collaborative), and first strand cDNAs were synthesized by incubating them for 1 hour at 12°C and overnight at 42°C with Super Script II (Invitrogen) and 10 pm of oligo-dT primer (5'-GCGGCTGAAGACGGCCTATGTGGCCTCCCTCGCGCCATCAG-T₁₇-3'). Template RNAs were then degraded by alkaline treatment and one fifth of the first strand

cDNAs were used as PCR templates. The PCR was carried out using the Gene Amp PCR kit (PerkinElmer) and the primers 5'-AGCATCGAGTCGGCCTTGTTG-3' and 5'-GCGGCTGAAGACGGCCTATGT-3' for 15 cycles of 1 minute at 94°C, 1 minute at 56°C and 2 minutes at 72°C. The resulting PCR fragments were size fractionated by a 12% polyacrylamide gel electrophoresis and the fraction corresponding to 150-250 bps was recovered. The quality and quantity of the obtained single-stranded first strand cDNAs were assessed using BioAnalyzer (Agilent). The sequencing of one nanogram of the size fractionated cDNA with a 454 DNA sequencer (Roche) was performed according to the manufacturer's instructions.

2.2 5'-EST classification

We could map 2,734 of the 2,859 5'-ESTs to the second version of the *C. intestinalis* genome (Dehal et al., 2002) using the blat (Kent, 2002) information from UCSC (large set) or generating our own (small set). As *trans*-splicing is characterized by the splicing of the SL to the 5'-end of a precursor mRNA, we looked for the presence of the known *C. intestinalis* SL sequence 5'-ATTCTATTTGAATAAG-3' (Vandenberghe et al., 2001) at the 5'-end of the 5'-ESTs. 614 5'-ESTs starting with the exact SL sequence were found. 67 5'-ESTs with small variations were also identified (Supplemental table 1). To confirm that these 681 5'-ESTs

were indeed *trans*-spliced, we checked whether or not the sequence directly downstream of the SL was mapped to the *C. intestinalis*. The 669 5'-ESTs for which it was the case were classified as SL(+) 5'-ESTs. The remaining 12 SL-containing 5'-ESTs were discarded because it was not possible to ascertain that they were genuine *trans*-spliced 5'-ESTs. The 2,053 5'-ESTs that did not contain the SL at their 5'-end were categorized as SL(-) 5'-ESTs.

2.3 Determination of genes expressed by *trans*-spliced ESTs

The *C. intestinalis* UniGene build #19 was used to determine which genes are expressed by *trans*-spliced ESTs (Supplemental data 2). 644 SL(+) 5'-ESTs clustered with ESTs representing 359 UniGene genes, and 1,952 SL(-) 5'-ESTs clustered with ESTs representing 446 UniGene genes. 41 UniGene genes were found in both groups. Because it is not possible to assign them either to the *trans*-spliced or non-*trans*-spliced gene group, these 41 UniGene genes were removed from both the SL(+) and the SL(-) UniGene genes groups. As a result, we obtained 318 SL(+) and 405 SL(-) UniGene genes associated with 423 SL(+) and 1,598 SL(-) 5'-ESTs.

2.4 Estimation of gene expression for SL(+) and SL(-) genes

Because of the correlation between the number of ESTs and the gene expression level (Adams et al., 1991; Wilcox et al., 1991), we used the 38,732 and 145,375 ESTs from UniGene

corresponding respectively to the 318 SL(+) and 405 SL(-) UniGene genes in order to investigate if SL(+) and SL(-) genes are expressed differently with regards to tissue or developmental stage. As these ESTs are derived from 26 libraries for which either the original tissue or the original developmental stage is known (Supplemental table 2), we could classify them into seven tissues and four developmental stages. Additionally, the ESTs from “Whole body” and “Adult” were used as reference for the analysis of tissues and developmental stages biases respectively.

2.5 Functional preference analysis

Because of the lack of functional annotations for *C. intestinalis* UniGene genes, we estimated the function of SL(+) and SL(-) *C. intestinalis* UniGene genes from the annotations of their homologous genes in other species. For each *C. intestinalis* UniGene gene, UniGene provides homology information with *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* and *Saccharomyces cerevisiae*. Table 1 indicates how many SL(+) and SL(-) *C. intestinalis* UniGene genes have homologs in each of these organisms. In total 261 SL(+) and 298 SL(-) *C. intestinalis* UniGene genes have homologs in at least one organism. Using the 'gene2accession' and 'gene2go' tables from NCBI, we could associate Gene Ontology (GO) IDs to the homologous genes. We used the generic GO slim mapping from the Gene Ontology Consortium

(www.geneontology.org) to determine which terms are associated preferentially with SL(+) or SL(-) genes (Supplementary data 3). The statistical significance was calculated using Fisher's exact test and Bonferroni's correction for multiple testing.

3. Results

3.1 Gene expression analysis

To determine which genes are expressed by *trans*-spliced ESTs, we clustered the 5'-ESTs with all the ESTs used in the *C. intestinalis* UniGene build #19. Although later versions of the *C. intestinalis* UniGene are available, we used this build because it contained 15,579 UniGene genes, which is close to the estimated number of *C. intestinalis* genes (15,852), whereas later builds only provide about 3,400 entries. Through clustering, 318 UniGene genes were classified as SL(+), and 405 as SL(-). These genes are respectively represented by 38,732 and 145,375 of the 653,547 ESTs used by UniGene. The 4.6% of the UniGene genes that were classified are therefore represented by 28.2% of the ESTs. Because the number of ESTs is correlated to the gene expression level (Adams et al., 1991; Wilcox et al., 1991), this observation indicates that the 5'-ESTs used to classify the UniGene genes are biased towards genes with higher expression levels. This bias is however not the same for SL(+) and SL(-) UniGene genes. Indeed, SL(+) UniGene genes, which correspond to 2.0% of the total UniGene gene population, are represented by 5.9% of the ESTs, while SL(-) UniGene genes (2.6% of the UniGene genes) are represented by 22.2% of the ESTs. The SL(-) UniGene genes are therefore more strongly biased towards highly expressed genes than the SL(+) UniGene genes. This is clearly visible in figure 1, which shows the distributions of the

number of ESTs per SL(+) and SL(-) UniGene genes, as well as for all UniGene genes. In this figure, the peaks indicate gene expression levels at which comparatively large number of genes are expressed. In figure 1A, we observe a single peak around 10^2 for SL(+) UniGene genes. The distribution of the number of ESTs per SL(-) UniGene genes shows however two peaks located slightly before 10^2 and 10^3 (figure 1B).

3.2 Preferential expression in tissues

The ESTs used by UniGene are derived from 26 libraries. For some of them, the tissue used for the experimental work is known. It is thus possible to annotate the UniGene genes categorized as SL(+) or SL(-) with one of seven tissues (“gonad”, “blood”, “neural complex”, “digestive gland”, “heart”, “testis”, “endostyle”), or with “whole body”. Figure 2 shows the proportion of SL(+) and SL(-) UniGene genes and associated ESTs for each tissue in adults. Comparing the number of SL(-) per SL(+) UniGene genes in each tissue with the one from the “whole body” category, a significant difference can only be observed for “endostyle”, where a preference for SL(-) genes is found (p-value of 3.34×10^{-6}). The situation is however different with regards to the associated ESTs, with significant differences being found for almost every tissue (table 2). In “gonad” and “blood” the number of SL(-) per SL(+) ESTs is of 2.53 and 3.19 respectively, which is lower than the value of 3.97 calculated for “whole

body”, while in “neural complex” (4.09) this number is not significantly different. In the remaining tissues, the ratio found is higher than in “whole body”, with the highest value found in “endostyle” (19.53). This indicates that depending on the tissue, SL(+) or SL(-) genes are preferentially expressed. Interestingly, the gender-related tissues “testis” and “gonad” are found at opposite ends of the preference range, with ratios of 5.77 and 2.53 respectively, despite testis being a part of gonads along with ovaries.

GO terms associated with homologs of SL(+) and SL(-) UniGene genes in seven organisms were used to identify possible functional biases in each tissue for either of the two populations. Table 3 summarizes the observed significant biases. It shows that, as reported previously (Satou et al., 2006), GO terms linked to the ribosome and its function from most organisms are associated preferentially with the SL(-) UniGene gene group in all tissues but “endostyle”. Two factors appear to be involved in the absence of significant preferences in “endostyle”: the small size of the library, 859 genes, and its bias towards SL(-) UniGene genes. A similarly strong association is not found for the SL(+) UniGene gene group, although the GO terms “protein transport”, “mitochondion”, and “cell cycle” are episodically found.

3.3 Preferential expression in developmental stages

Similarly as with tissues, it is possible to annotate the UniGene genes categorized as SL(+) or SL(-) with one of four developmental stages (“egg”, “embryo”, “larval”, “juvenile”), or with “adult”. Figure 3 shows the proportion of SL(+) and SL(-) UniGene genes and associated ESTs from whole body samples for each developmental stage. While no significant differences are observed when comparing the number of SL(-) per SL(+) UniGene genes in each developmental stage with the one from the “adult” category, only significant differences are found in the case of the associated ESTs (table 4). The number of SL(-) per SL(+) ESTs observed in “egg” and “embryo”, 1.68 and 2.53 respectively, is lower than the value of 3.97 calculated for “adult”, while the ratios found in “larval” (5.76) and “juvenile” (7.55) are higher.

The significant biases observed for the SL(+) or the SL(-) UniGene genes population at each developmental stage by using the GO terms associated with homologs of SL(+) and SL(-) UniGene genes in seven organisms are summarized in table 5. As with tissues, for most organisms, GO terms linked to the ribosome and its function are associated preferentially with the SL(-) UniGene gene group. At every developmental stage, as well as in adults, the GO term “mitochondrion” is found to be preferentially associated with the SL(+) UniGene gene population when using *Saccharomyces cerevisiae* homologs. The GO terms “protein transport” and “cell cycle” are again also observed episodically.

4. Discussion

We used two 5'-ESTs datasets from different origins. The 5'-ESTs of the large set were obtained from a mixture of egg, tailbud embryo, larva and young adults, while the 5'-ESTs of the small set were obtained from an adult individual. Satou et al. reported that the large dataset is not biased towards either SL(-) or SL(+) when compared with conventional ESTs from cDNA libraries representing the same developmental stages (Satou et al., 2006).

When considering only the large set, 299 SL(-) and 262 SL(+) UniGene genes can be obtained. 15,772 and 5,138 UniGene ESTs from the whole body of adults correspond respectively to the SL(-) and the SL(+) UniGene genes, and constitute an unbiased set of conventional ESTs to evaluate the small set of 5'-ESTs. From the small set, 518 and 142 5'-ESTs could be clustered with 205 SL(-) and 90 SL(+) UniGene genes respectively. According to Fisher's exact test, the number of SL(-) per SL(+) 5'-EST (518/142) is not significantly different from that of the conventional ESTs from the whole body of adults obtained using the large set (15,772/5,138) (p-value of 0,073). The small set is therefore unlikely to be biased for or against the *trans*-spliced mRNA subpopulation.

Additionally, although embryonic and adult mRNA populations may be different, both can be used to identify *trans*-spliced and non-*trans*-spliced genes. There is indeed no evidence that a *trans*-spliced gene becomes non-*trans*-spliced, or that a non-*trans*-spliced gene becomes

trans-spliced during development, and it can thus be assumed that if such an event happens, it is very rare. This and the fact the proportion of *trans*-spliced and non-*trans*-spliced genes in tissues or developmental stages remains the same, with the exception of “endostyle”, indicates that the origin of the 5'-EST sampling is unlikely to influence the observed differences in gene expression (tables 2 and 4). The question of a possible bias of the libraries used remains however to be more thoroughly investigated. It should be noted though that most of these libraries originate from one group, and thus are likely to be biased in the same way, if they are. In that case, despite the influence of this possible bias on the presented results, the observed differences should remain valid.

The Gene Ontology website (www.geneontology.org) reports 4,324 gene products annotated with “structural molecule activity”. Among them, 2,674 gene products (61.8%) are annotated at a lower level with “structural constituent of ribosome”. This preponderance of genes annotated with “structural constituent of ribosome” within the genes annotated with “structural molecule activity” provide an acceptable explanation to why the GO term “structural molecule activity” is associated preferentially with the SL(-) UniGene gene population, along with the expected GO terms “ribosome” and “translation”.

From our analysis of preferential expression in tissues, we found a strong bias for SL(+) associated ESTs in “gonad”, and a strong bias for SL(-) associated ESTs in “testis”. As “gonad” includes both “testis” (male) and “ovary” (female), this implies that “ovary” must be

even more biased toward SL(+) associated ESTs than “gonad”. Provided this is correct, it would explain the abundance of SL(+) associated ESTs we also observed in early developmental stages, when most of the expressed genes are of maternal origin. Indeed, the number of SL(-) per SL(+) associated ESTs appears to increase during development, starting at 1.68 in “egg” and finishing at 7.55 in “juvenile”.

It is furthermore interesting to note that Azumi et al. reported that many of the genes contained in the maternal cluster of their study encoded proteins functioning in mitochondria and the cell cycle (Azumi et al., 2007), and that although we did not find any GO term strongly bias towards SL(+) UniGene genes, the GO terms “mitochondrion”, “protein transport” and “cell cycle” were episodically observed. The studies of operon structures in *C. elegans* and *C. intestinalis* bring further support to the possible association of mitochondrion and *trans*-splicing. From *C. elegans*, we learned that the genes most frequently included in operons are those that encode mitochondrial proteins and basic machinery of gene expression, and that many of the genes observed in operon structures are subject to *trans*-splicing (Blumenthal, 2005). Several examples of *C. intestinalis* operons together with evidences of the involvement of *trans*-splicing in the resolution of their inside genes were given by Satou et al. (Satou et al., 2006). In fact 19 of these inside genes were reported to be associated with the *trans*-splicing of 26 5'-ESTs. Among them, we observed 3 mitochondrial related, 2 transcription related, 1 splicing related, and 1 translation related genes; an observation that is

consistent with the one made in *C. elegans*.

While clustering the 5'-ESTs to all the UniGene ESTs in order to identifying SL(+) and SL(-) UniGene genes we found 41 UniGene genes to which both *trans*-spliced and non-*trans*-spliced 5'-ESTs were associated. For 13 of them, the 5'-ESTs were assigned to the same UniGene gene by clustering, but mapped by UCSC to different locations in the *C. intestinalis* genome. This discrepancy is mostly due to bad clustering, although bad mapping might also be involved. A closer investigation of the remaining 28 SL(+/-) UniGene genes was carried out (Supplemental figure 1). Among them, we found 9 cases possibly due to truncation of *trans*-spliced ESTs, 4 cases resulting from erroneous gene models and truncation, and 2 cases where the SL is apparently truncated. Finally, we observed 13 cases where the *trans*-splicing or the lack of *trans*-splicing could be considered as an alternative splicing variant.

Because the majority of the possible truncation cases are due to only one 5'-EST being classified as non-*trans*-spliced, we investigated how truncation could influence the results presented here by either removing all SL(-) UniGene genes for which only one 5'-EST was available, or by changing it to SL(+). Both of these actions are too extreme, as it is unlikely that all these 5'-ESTs are truncated. Nevertheless, we recalculating the significance of the differences in the number of SL(-) per SL(+) genes or ESTs in tables 2 and 4, we could not observe any change susceptible to invalidate our conclusions (data not shown).

Although our analysis is based on a limited amount of 5'-ESTs representing relatively strongly expressed genes, we observed for the first time that the expression levels of *trans*-spliced and non-*trans*-spliced genes vary differently during development and across tissues. Future work using a comprehensive set of 5'-ESTs representing accurately the *C. intestinalis* gene population should however be carried out to determine whether or not this observation remains valid for genes with lower expression levels.

Acknowledgement

We thank Professor Kusakabe for valuable comments. Computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo. This work was supported by Culture of Japan and BIRD of Japan Science and Technology Agency (JST).

References

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merrill, C.R., Wu, A., Olde, B., Moreno, R.F. and et al., 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651-1656.

- Azumi, K., Sabau, S.V., Fujie, M., Usami, T., Koyanagi, R., Kawashima, T., Fujiwara, S., Ogasawara, M., Satake, M., Nonaka, M., Wang, H.G., Satou, Y. and Satoh, N., 2007. Gene expression profile during the life cycle of the urochordate *Ciona intestinalis*. Dev. Biol. 308, 572-582.
- Blaxter, M. and Liu, L., 1996. Nematode spliced leaders--ubiquity, evolution and utility. Int. J. Parasitol. 26, 1025-1033.
- Blumenthal, T., 2005. Trans-splicing and operons, WormBook. The *C. elegans* Research Community, doi/10.1895/wormbook.1.5.1, <http://www.wormbook.org>.
- Blumenthal, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W.L., Duke, K., Kiraly, M. and Kim, S.K., 2002. A global analysis of *Caenorhabditis elegans* operons. Nature 417, 851-854.
- Blumenthal, T. and Gleason, K.S., 2003. *Caenorhabditis elegans* operons: form and function. Nat. Rev. Genet. 4, 112-120.
- Bonen, L., 1993. Trans-splicing of pre-mRNA in plants, animals, and protists. Faseb J. 7, 40-46.
- Campbell, D.A., Thomas, S. and Sturm, N.R., 2003. Transcription in kinetoplastid protozoa: why be normal? Microbes Infect. 5, 1231-1240.
- Conklin, E.G., 1905. The organisation and cell lineage of the ascidian egg. J. Acad. Natl. Sci. 13, 1-119.

- Corbo, J.C., Di Gregorio, A. and Levine, M., 2001. The ascidian as a model organism in developmental and evolutionary biology. *Cell* 106, 535-538.
- Davis, R.E., 1996. Spliced leader RNA trans-splicing in metazoa. *Parasitol. Today* 12, 33-40.
- Davis, R.E. and Hodgson, S., 1997. Gene linkage and steady state RNAs suggest trans-splicing may be associated with a polycistronic transcript in *Schistosoma mansoni*. *Mol. Biochem. Parasitol.* 89, 25-39.
- Dehal, P., et al., 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* 298, 2157-2167.
- Ganot, P., Kallesoe, T., Reinhardt, R., Chourrout, D. and Thompson, E.M., 2004. Spliced-leader RNA trans splicing in a chordate, *Oikopleura dioica*, with a compact genome. *Mol. Cell. Biol.* 24, 7795-7805.
- Gunzl, A., Bruderer, T., Laufer, G., Schimanski, B., Tu, L.C., Chung, H.M., Lee, P.T. and Lee, M.G., 2003. RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. *Eukaryot. Cell* 2, 542-551.
- Hastings, K.E., 2005. SL trans-splicing: easy come or easy go? *Trends Genet.* 21, 240-247.
- Kent, W.J., 2002. BLAT--the BLAST-like alignment tool. *Genome Res* 12, 656-664.
- Krause, M. and Hirsh, D., 1987. A trans-spliced leader sequence on actin mRNA in *C. elegans*. *Cell* 49, 753-761.
- Laird, P.W., 1989. Trans splicing in trypanosomes--archaism or adaptation? *Trends Genet.* 5,

204-208.

- Lall, S., Friedman, C.C., Jankowska-Anyszka, M., Stepinski, J., Darzynkiewicz, E. and Davis, R.E., 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.
- Lee, M.G. and Van der Ploeg, L.H., 1997. Transcription of protein-coding genes in trypanosomes by RNA polymerase I. Annu. Rev. Microbiol. 51, 463-489.
- Maroney, P.A., Denker, J.A., Darzynkiewicz, E., Laneve, R. and Nilsen, T.W., 1995. Most mRNAs in the nematode *Ascaris lumbricoides* are trans-spliced: a role for spliced leader addition in translational efficiency. RNA 1, 714-723.
- Maruyama, K. and Sugano, S., 1994. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene 138, 171-174.
- Nilsen, T.W., 1995. trans-splicing: an update. Mol. Biochem. Parasitol. 73, 1-6.
- Nishida, H., 1987. Cell lineage analysis in ascidian embryos by intracellular injection of a tracer enzyme. III. Up to the tissue restricted stage. Dev Biol 121, 526-541.
- Passamaneck, Y.J. and Di Gregorio, A., 2005. *Ciona intestinalis*: chordate development made simple. Dev. Dyn. 233, 1-19.
- Pouchkina-Stantcheva, N.N. and Tunnacliffe, A., 2005. Spliced leader RNA-mediated trans-splicing in phylum Rotifera. Mol. Biol. Evol. 22, 1482-1489.

- Satoh, N., 2003. The ascidian tadpole larva: comparative molecular development and genomics. *Nat. Rev. Genet.* 4, 285-295.
- Satoh, N. and Jeffery, W.R., 1995. Chasing tails in ascidians: developmental insights into the origin and evolution of chordates. *Trends Genet.* 11, 354-359.
- Satoh, N., Satou, Y., Davidson, B. and Levine, M., 2003. *Ciona intestinalis*: an emerging model for whole-genome analyses. *Trends Genet.* 19, 376-381.
- Satou, Y., Hamaguchi, M., Takeuchi, K., Hastings, K.E. and Satoh, N., 2006. Genomic overview of mRNA 5'-leader trans-splicing in the ascidian *Ciona intestinalis*. *Nucleic Acids Res.* 34, 3378-3388.
- Satou, Y., et al., 2001. Gene expression profiles in *Ciona intestinalis* tailbud embryos. *Development* 128, 2893-2904.
- Stover, N.A. and Steele, R.E., 2001. Trans-spliced leader addition to mRNAs in a cnidarian. *Proc. Natl. Acad. Sci. USA* 98, 5693-5698.
- Sutton, R.E. and Boothroyd, J.C., 1986. Evidence for trans splicing in trypanosomes. *Cell* 47, 527-535.
- Vandenberghe, A.E., Meedel, T.H. and Hastings, K.E., 2001. mRNA 5'-leader trans-splicing in the chordates. *Genes Dev.* 15, 294-303.
- Wilcox, A.S., Khan, A.S., Hopkins, J.A. and Sikela, J.M., 1991. Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to

STSs: implications for an expression map of the genome. *Nucleic Acids Res* 19, 1837-1843.

Yuasa, H.J., Kawamura, K., Yamamoto, H. and Takagi, T., 2002. The structural organization of ascidian *Halocynthia roretzi* troponin I genes. *J. Biochem.* 132, 135-141.

Zeiner, G.M., Sturm, N.R. and Campbell, D.A., 2003. The *Leishmania tarentolae* spliced leader contains determinants for association with polysomes. *J. Biol. Chem.* 278, 38269-38275.

Tables and figures

Table 1

Homologous genes between *C. intestinalis* and seven other organisms.

Organisms	#Homologous Gene	
	SL(+)	SL(-)
<i>C. intestinalis</i>	261 (228)	298 (272)
<i>H. sapiens</i>	242 (109)	267 (137)
<i>M. musculus</i>	182 (104)	229 (125)
<i>R. norvegicus</i>	118 (66)	194 (105)
<i>C. elegans</i>	211 (109)	240 (124)
<i>A. thaliana</i>	172 (100)	194 (138)
<i>D. melanogaster</i>	116 (66)	158 (95)
<i>S. cerevisiae</i>	143 (44)	166 (62)

For *C. intestinalis*, the number of genes with homologs in any of the seven organisms is given; otherwise, the numbers of *C. intestinalis* genes with homologs in the corresponding organism is given. The number of *C. intestinalis* genes for which GO annotations from homologs are available is given in parenthesis.

Table 2

Tissue specificity of *trans*-splicing

	Genes				ESTs			
	SL(-)	SL(+)	SL(-) / SL(+)	p-value	SL(-)	SL(+)	SL(-) / SL(+)	p-value
Whole body	311	269	1.16		17230	4340	3.97	
Gonad	247	212	1.17	1	7180	2842	2.53	4.30×10^{-57}
Blood	248	220	1.13	1	8767	2747	3.19	3.18×10^{-14}
Neural complex	198	165	1.20	1	3859	944	4.09	1
Digestive gland	247	212	1.17	1	10453	1932	5.41	1.32×10^{-24}
Heart	204	164	1.24	1	5098	932	5.47	7.80×10^{-16}
Testis	160	131	1.22	1	2614	453	5.77	4.89×10^{-12}
Endostyle	102	30	3.40	3.34×10^{-6}	1445	74	19.53	8.65×10^{-62}

Statistical significance values were obtained using Fisher's exact test and corrected with Bonferroni's method. SL(+) ESTs and SL(-) ESTs indicate the number of conventional ESTs from adults corresponding to SL(+) genes and SL(-) genes, respectively, in each tissue.

Table 3

GO term preference in tissues

GO term	Preference	En.	Bl.	He.	Ne.	Go.	Te.	Di.	Wh.	Un.
Structural molecule activity	SL(-)	0	6	5	6	6	7	6	7	7
Translation	SL(-)	0	5	6	5	5	5	6	5	4
Ribosome	SL(-)	0	6	5	5	5	5	4	6	5
RNA binding	SL(-)	0	1	1	2	1	1	1	1	2
Intracellular	SL(-)	1	1	1	1	1	1	1	1	1
Nucleic acid binding	SL(-)	0	1	1	1	1	1	1	1	1
Protein transport	SL(+)	0	1	1	0	0	1	0	1	0
Mitochondrion	SL(+)	0	1	0	0	0	0	1	1	1
Cell cycle	SL(+)	0	0	0	0	1	1	0	0	0
Receptor activity	SL(+)	0	1	0	0	0	0	0	0	0
Response to abiotic stimulus	SL(+)	0	0	0	0	0	0	0	1	0
Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	SL(+)	0	0	1	0	0	0	0	0	0

The number of organisms from which homology information resulted in a statistically significant bias towards SL(-) or SL(+) genes for each GO term is given. En.: endostyle; Bl.: blood; He.: heart; Ne.: neural complex; Go.: gonad; Te.: testis; Di.: digestive gland; Wh.: whole body; Un.: unknown tissue.

Table 4

Developmental stage specificity of *trans*-splicing

	Genes				ESTs			
	SL(-)	SL(+)	SL(-) / SL(+)	p-value	SL(-)	SL(+)	SL(-) / SL(+)	p-value
Adult	311	269	1.16		17230	4340	3.97	
Egg	269	259	1.04	1	6896	4115	1.68	3.08×10^{-239}
Embryo	340	298	1.14	1	33352	13157	2.53	1.05×10^{-116}
Larval	267	211	1.27	1	14603	2536	5.76	4.74×10^{-42}
Juvenile	330	278	1.19	1	32681	4328	7.55	2.36×10^{-163}

Statistical significance values were obtained using Fisher's exact test and corrected with Bonferroni's method. SL(+) ESTs and SL(-) ESTs indicate the number of conventional ESTs from whole body samples corresponding to SL(+) genes and SL(-) genes, respectively, at each developmental stage.

Table 5

GO term preference in developmental stages

GO term	Preference	Eg.	Em.	La.	Ju.	Ad.	Un.
Structural molecule activity	SL(-)	7	7	7	7	7	7
Ribosome	SL(-)	5	6	6	6	6	5
Translation	SL(-)	5	5	5	5	5	4
RNA binding	SL(-)	1	2	1	1	1	2
Nucleic acid binding	SL(-)	1	1	1	1	1	1
Mitochondrion	SL(+)	1	1	1	1	1	1
Intracellular	SL(-)	1	1	0	1	1	1
Protein transport	SL(+)	1	0	1	1	0	0
Response to abiotic stimulus	SL(+)	0	1	0	0	1	0
Cell cycle	SL(+)	0	0	1	0	0	0
Metabolic process	SL(+)	0	0	0	1	0	0

The number of organisms from which homology information resulted in a statistically significant bias towards SL(-) or SL(+) genes for each GO term is given. Eg.: egg; Em.: embryo; La.: larval; Ju.: juvenile; Ad.: adult; Un.: unknown developmental stage.

Figure 1

Distribution of EST numbers per SL(+) and SL(-) UniGene gene.

The horizontal axis shows the logarithm in base 10 of the number of ESTs clustered in one gene, and the vertical axis shows the number of genes. The upper graph is from the SL(+) gene group, and the lower graph is from the SL(-) gene group.

Figure 2

Preferential expression in tissues

The bars give the proportions of SL(+) (light gray) and SL(-) (dark gray) ESTs from adult samples observed in each tissue. The thin line indicate the proportion of SL(+) and SL(-) UniGene genes. Actual numbers can be found in table 2.

Figure 3

Preferential expression in developmental stages

The bars give the proportions of SL(+) (light gray) and SL(-) (dark gray) ESTs from whole body samples observed at each developmental stage. The thin line indicate the proportion of SL(+) and SL(-) UniGene genes. Actual numbers can be found in table 4.

Figure 1

Figure 2

Figure 3