3378–3388 *Nucleic Acids Research, 2006, Vol. 34, No. 11*
doi:10.1093/nar/gkl418

Genomic overview of mRNA 5′-leader *trans*-splicing in the ascidian *Ciona intestinalis*

Yutaka Satou[1,*], Makoto Hamaguchi[1], Keisuke Takeuchi[1], Kenneth E. M. Hastings[2] and Nori Satoh[1,3]

[1]Department of Zoology, Graduate School of Science, Kyoto University, Sakyo, Kyoto 606-8502, Japan, [2]Montreal Neurological Institute and Department of Biology, McGill University, 3801 University St. Montreal, Quebec, Canada H3A 2B4 and [3]CREST, Japan Science Technology Agency, Kawaguchi, Saitama, 330-0012, Japan

Received March 25, 2006; Revised and Accepted May 19, 2006

ABSTRACT

Although spliced leader (SL) *trans*-splicing in the chordates was discovered in the tunicate *Ciona intestinalis* there has been no genomic overview analysis of the extent of *trans*-splicing or the make-up of the *trans*-spliced and non-*trans*-spliced gene populations of this model organism. Here we report such an analysis for *Ciona* based on the oligo-capping full-length cDNA approach. We randomly sampled 2078 5′-full-length ESTs representing 668 genes, or 4.2% of the entire genome. Our results indicate that *Ciona* contains a single major SL, which is efficiently *trans*-spliced to mRNAs transcribed from a specific set of genes representing ~50% of the total number of expressed genes, and that individual *trans*-spliced mRNA species are, on average, 2–3-fold less abundant than non-*trans*-spliced mRNA species. Our results also identify a relationship between *trans*-splicing status and gene functional classification; ribosomal protein genes fall predominantly into the non-*trans*-spliced category. In addition, our data provide the first evidence for the occurrence of polycistronic transcription in *Ciona*. An interesting feature of the *Ciona* polycistronic transcription units is that the great majority entirely lack intercistronic sequences.

INTRODUCTION

The ascidian tunicate *Ciona intestinalis* is a chordate whose 160 Mb genome, with ∼16 000 genes, is considerably simpler than those of vertebrates, such as man, mouse and pufferfish, which contain ∼30 000 genes (1–3). The smaller number of genes, comparatively short intergenic distances, and the robust and experimentally accessible nature of its early development have made the ascidian an important model system for genetic analysis of chordate development (4,5). In-depth knowledge of the ascidian genome will also contribute to our understanding of chordate evolution and the origins of the vertebrate genome. This developmental and evolutionary relevance has driven extensive molecular genetic studies so that *Ciona* has become one of the better-characterized animals in terms of genomics resources (6,7).

Recent studies have uncovered an unexpected and striking genomic difference between tunicates and vertebrates. This difference concerns mRNA 5′-leader *trans*-splicing, or spliced leader (SL) *trans*-splicing. In SL *trans*-splicing, the original 5′ ends of some pre-mRNAs are discarded and are replaced, in a spliceosomal mechanism, by the 5′-region of a small, specialized donor RNA, the SL RNA (8). Because multiple pre-mRNAs are *trans*-spliced by the same SL RNA species, SL sequences are found as common sequences at the 5′ ends of diverse mRNA species. SL *trans*-splicing occurs in tunicates, including *Ciona* (9,10). However, despite intensive genetics research, it has not been observed in any vertebrate, and presumably does not occur in that group. From its patchy distribution among the eukaryotic kingdoms and phyla, it is not clear if SL *trans*-splicing is an ancestral eukaryotic mechanism that has been secondarily lost in several lineages, e.g. the vertebrates, arthropods, plants, fungi, or if it was absent from the ancestral eukaryote and arose independently within each of the several phyla in which it is now known to occur: nematodes, flatworms, chordates, cnidarians, rotifers and protist euglenozoans (11–14). In-depth genomic studies of SL *trans*-splicing organisms, e.g. tunicates, and of related non-*trans*-splicing organisms, e.g. vertebrates, are likely to generate insight into the evolution of SL *trans*-splicing, and the implications of its evolutionary gain or loss for other aspects of genome organization and function.

The functions of SL *trans*-splicing are partly, but not entirely, understood. Its best-known role is to resolve polycistronic transcripts into individual 5′-capped monocistronic mRNAs, a process that has been extensively studied in nematodes (15,16) flatworms (17) and in euglenozoan protists

[*]To whom correspondence should be addressed. Tel: +81-75-753-4095; Fax: +81-75-705-1113; Email: yutaka@ascidian.zool.kyoto-u.ac.jp

including kinetoplastids [trypanosomes, where it is the dominant mechanism of gene expression (18–22)]. Most known SL-resolved polycistronic transcription units, or operons, include short intercistronic sequences that in the nematode *Caenorhabditis* include *cis*-elements playing an active role in directing *trans*-splicing of downstream cistrons (23). In *Caenorhabditis*, genes within operons are in some cases functionally related (16), and have a significant overall tendency to show similar patterns of mRNA accumulation (24). Thus operons could represent a mechanism for coordinating gene expression. However, because many operons contain genes that have no obvious functional relationship and/or are not coordinately expressed (16), additional non-specific factors, such as genome compaction may also contribute to operon evolution. Apart from its role in operons it is likely that SL-*trans*-splicing has additional functions, e.g. in regulating mRNA stability or translation (25,26) or in removing potentially deleterious sequences from pre-mRNA 5′-untranslated regions (5′-UTR) (27), because in nematodes and flatworms, and possibly all *trans*-splicing metazoa, only a modest fraction of *trans*-spliced genes are in operons; the majority are transcribed mono-cistronically.

SL-*trans*-splicing occurs on mono- or poly-cistronic pre-mRNA targets only at splice accepter sites that do not have a partner donor site upstream in the transcript, i.e. unpaired acceptor sites. Acceptor sites that are paired with an upstream partner donor site preferentially undergo *cis*-splicing with removal of the intervening intron, rather than *trans*-splicing (27). All known *trans*-splicing metazoa appear to have a significant class of conventionally-expressed genes that do not contain unpaired acceptor sites and hence do not undergo *trans*-splicing. However, there have been no sequence-based overview studies of both the *trans*-spliced and non-*trans*-spliced gene classes in any metazoan and the biological implications of the division of the genome into these two major gene classes remain unexplored.

The first report of SL *trans*-splicing in the chordates identified a 16 nt *trans*-spliced leader transferred to at least seven mRNA species in *Ciona* (9). A distinct 40 nt SL was subsequently reported in another tunicate, *Oikopleura dioica*, which belongs to the distantly-related class Appendicularia (*Larvaceae*), whose morphology, behavior and developmental and ecological strategies differ markedly from ascidians, and whose genome evolution has featured a marked overall compaction (10). In *Oikopleura* 12–24% of genes give rise to *trans*-spliced mRNAs, including some genes in SL-resolved operons (10). In *Ciona* the overall extent of *trans*-splicing is unknown, and polycistronic transcription has not been reported.

In order to advance our understanding of the *Ciona* genome in relation to SL-*trans*-splicing, we have carried out a global overview analysis of the *trans*-spliced and non-*trans*-spliced gene populations. Our goals were to answer the following questions: What fraction of *Ciona* genes gives rise to *trans*-spliced mRNAs? Are the *trans*-spliced and non-*trans*-spliced gene classes specialized in terms of gene function? Does the *Ciona* genome contain operons that are resolved by SL *trans*-splicing? Is the currently-known 16 nt SL the only one, or are additional, novel, SL sequences also used, as in *Caenorhabditis*, which contains a second SL RNA, SL2, devoted to polycistron resolution? Our study reports the first broad samplings of both the *trans*-spliced and non-*trans*-spliced mRNA subpopulations of any organism, and reveals the differential distribution of a functional gene class (ribosomal protein genes) between these mRNA populations. It also provides the first evidence for polycistronic transcription in *Ciona*. Moreover our results reveal an interesting feature of polycistronic transcription in *Ciona*, i.e. a predominance of operons entirely lacking intercistronic sequences.

MATERIALS AND METHODS

Ascidian eggs and embryos

Ciona intestinalis adults were cultivated at the Maizuru Fisheries Research Station of Kyoto University, Maizuru city, facing the Sea of Japan. They were maintained in aquaria in our laboratory at Kyoto University under constant light to induce oocyte maturation. Eggs and sperm were obtained surgically from gonoducts. After fertilization, embryos were reared at ∼18°C in Millipore-filtered seawater containing 50 μg/ml streptomycin sulfate.

Construction of a full-length enriched cDNA library

Total RNA was isolated independently from four different developmental stages (eggs, tailbud embryos, larvae and young adults) of *Ciona intestinalis* by the acid guanidinium thiocyanate–phenol–chloroform method (28). Oligo-capping of a mixture of equal amounts of isolated RNAs was performed using a commercially available kit (GeneRacer kit, Invitrogen). Oligo-capped RNA was reverse-transcribed with tagged-oligo-(dT) primer, and the resultant cDNA was amplified by 15 cycles of PCRs with Pfu DNA polymerase using primers for the capping-oligo and the tag in the oligo-(dT) primer. The amplified cDNAs were size-fractionated by gel chromatography. After treating with *Taq* DNA polymerase for 5 min at 72°C, the cDNA was cloned into pGEM-T vector (Promega).

5′ End sequencing of oligo-capped cDNA clones

cDNA inserts were PCR amplified using M13 reverse and forward primers. Successful amplifications were confirmed by agarose gel electrophoresis. After purification of the PCR products with Montage-PCR Filter Units (Millipore), their sequences were determined by conventional procedures using the big-dye terminator kits on an ABI PRISM 3700 DNA Analyzer (Applied Biosystems), and the same primers used for amplification.

Informatics analyses

5′-Terminal sequences for oligo-capping cDNA clones, full-length enriched ESTs (simply termed full-length ESTs hereafter), were BLAST-searched against themselves, and the results were used for clustering with a threshold score of 150. The clustering result was evaluated based on mapping information of the full-length ESTs onto the genome. We found that one cluster contained eight different actin genes because of high conservation of their nucleotide sequences, and we manually corrected this problem. A unique number was assigned to each cluster. To find spliced-leaders, 5′ end sequences of length 20 nt were compared with each other using the CROSSMATCH program (29).

Analysis of polycistronic transcription units in the *Ciona* genome was based on a set of gene models, called Kyotograil2004, recently predicted by the grailexp program based on ~680 000 ESTs and ~6500 full insert cDNA sequences (30). Three independent approaches were used to uncover candidate operons: analysis of gene models upstream of mapped SL-full-length ESTs, genome-wide search for closely-spaced head-to-tail gene models, and search among conventional ESTs for dicistronic transcripts.

Gene models upstream of genome-mapped full-length ESTs: All full-length ESTs were aligned with the genome sequence using the BLAT program (31) and the genomic distance between the 5′ end of each full-length EST and the nearest end of the nearest upstream non-overlapping gene model was tabulated.

Genome search for closely-spaced head-to-tail gene models: All cases where non-overlapping neighbouring gene models in the same transcriptional orientation were separated by less than or equal to 100 nt were recovered. In some cases more than two gene models in a row satisfied the criteria. Each such group of two or more gene models was assigned a unique group identification number. In addition, groups described more in detail in the present paper were given an independent operon identification number. All gene groups were manually examined on the genome browser (30) on which full-length ESTs and an extensive collection of conventional ESTs were also mapped. Only groups in which both gene models were supported by ESTs and/or full-length ESTs were considered to be candidate operons.

Search for dicistronic conventional ESTs: All non-redundant conventional ESTs were mapped onto the genome by BLAT and coordinates were compared with a list of grailexp gene model coordinates. ESTs were recovered that mapped to two non-overlapping neighbouring gene models having the same transcriptional orientation.

GO assesment: The proteins deduced from the gene models were searched against human proteins with the BLAST program (32). The human protein set used was a group of 11 632 proteins annotated with GO terms in the molecular function category among the reference sequences in NCBI (release 9). The cut-off value for significant similarities was set to $e = 1E - 15$. GO terms in the molecular function category for each top-scoring hit were compared to determine whether genes within each pair or group were functionally related. GO terms in the fourth rank and their children were treated as their parent GO terms in the third rank for efficient comparisons. Among 179 GO terms in the third rank of the molecular function category, 42 GO terms were associated with genes having significant human/*Ciona* similarity.

Determination of operon intercistron boundaries: SL-full-length EST sequence data precisely localized the downstream gene's *trans*-splice acceptor site. The 3′ end, poly(A)-adjacent sequences, of mRNAs derived from upstream genes were identified through various means. Because our major conventional cDNA 3′ EST sequencing approach in prior studies (33) had been based on priming with an oligo(dT)-containing primer, most 3′ EST sequencing runs were lacking several bases immediately adjacent to the poly(A). However for some short-insert cDNAs, conventional 5′-EST sequences provided poly(A)-adjacent mRNA sequences. In addition, our collection of full-insert cDNAs (6), previously sequenced by primer walking and vector-based priming, also provided poly(A)-adjacent sequence for several mRNAs. Finally, where necessary for the present study, conventional EST clones were resequenced using a vector-based primer that permitted determination of poly(A)-adjacent mRNA sequence.

RESULTS

5′-ESTs from a full-length enriched cDNA library

In the initial discovery of SL *trans*-splicing in *Ciona*, seven *trans*-spliced mRNAs were identified but the overall genomic extent of *trans*-splicing was not established (9). In the present study, we used a full-length cDNA cloning/DNA sequencing approach to identify a significant and representative fraction of the mRNA species in the *trans*-spliced and non-*trans*-spliced subpopulations.

Trans-spliced and non-*trans*-spliced mRNAs can be recognized by the presence or absence of a 5′-SL. The presence/absence of the known *Ciona* 16 nt SL at mRNA 5′-termini cannot be determined from existing *Ciona* EST data because these ESTs are based on cDNA molecules produced by conventional cloning methods, which inevitably lose 8–21 nt of mRNA 5′-sequence information in the final double-stranded cDNA (34). In order to produce cDNA clones that do contain the extreme mRNA 5′-sequence, we used the oligo-capping method (35). To obtain a broad and representative gene sampling, we determined mRNA 5′-terminal sequences for 2078 randomly-picked oligo-capping cDNA clones generated from a mixture of egg, tailbud embryo, larva and young adult mRNA (DDBJ accession nos.: BW648671–BW650748). Within this study, we term these 5′ end sequences full-length ESTs to discriminate them from conventional ESTs, termed simply ESTs, in order to avoid confusion when we compare these two types of EST data, and in recognition of the unique genomic applications of 5′-complete, as opposed to 5′-incomplete, mRNA sequence data. Based on sequence similarities, the 2078 full-length ESTs were organized into 668 clusters of related clones, each cluster representing a different mRNA species/gene (Supplementary Table S1). Because 15 852 genes are predicted in the *Ciona* draft genome sequence (7), the number of genes covered by the full-length ESTs corresponds to 4.2% of the whole gene set. With a sample of this size the relative numbers of *trans*-spliced and non-*trans*-spliced genes in the genome can be estimated with the 5% confidence interval at the 99% confidence level. As discussed below, the evaluation of the full-length EST set by comparison with a non-biased conventional EST set, which we had obtained by independent experiments previously (6), suggested that the present full-length EST set does not contain a strong bias for or against the *trans*-spliced and non-*trans*-spliced mRNA subpopulation.

Ciona has only one major SL

To identify *trans*-spliced leaders as 5′-terminal sequences shared by diverse mRNA species, we cross-compared the first 20 nt of the full-length EST set by the CROSSMATCH

Table 1. SL sequence variants observed in SL-full-length ESTs

SL sequence	Number of full-length ESTs	Number of clusters/genes
ATTCTATTTGAATAAG	517	307
_TTCTATTTGAATAAG	14	14
ATTCTATTTAAATAAG	4	4
ATTTCTATTTGAATAAG	4	4
_CTATTTGAATAAG	3	3
AATTCTATTTGAATAAG	3	3
GTATTCTATTTGAATAAG	3	3
ATTCTAATTGAATAAG	3	2
_TCTATTTGAATAAG	2	2
ATTCTATTAGAATAAG	1	1
ATTCTA_TTGAATAAG	1	1
ATTCTATTTCAATAAG	1	1
ATTCTATTTGAAAAAG	1	1
ATTCTATTTGAACAAG	1	1
ATTCTATTTGAAGAAG	1	1
ATTCTATTTGA_____	1	1
ATTCTGTTTGAATAAG	1	1

program (29). This identified a single common 5′-sequence of 16 nt, identical to the previously-reported SL sequence (9) that was shared, with a low-level of microheterogeneity (Table 1), by 563 full-length ESTs (termed SL-full-length ESTs) representing 332 genes. Many *Ciona* SL genes were found in unassembled part of genome sequences, which were set aside from the main assembly because of high repetitiveness (7), and such microheterogeneity was actually found there (Supplementary Figure S1). Consistent with a *trans*-splicing origin for the SL in these SL-full-length EST mRNA sequences, the SL sequence itself was not present in the genomic DNA regions encoding the mRNAs. The 5′-sequences of the remaining 1515 full-length ESTs (non-SL-full-length ESTs), representing 350 genes, were unique in that each was associated with only one mRNA species (fourteen genes were represented by both SL-full-length ESTs and non-SL-full-length ESTs). The absence of additional shared 5′-sequences strongly suggests that *Ciona* has only one SL or that any additional SL that might exist could be associated with at most a very minor fraction of the mRNA population (the probability that an additional SL, associated with even as few as 1% of mRNA molecules, would not have been sampled twice or more in our dataset is <1%). In *Caenorhabditis* [and other Clade V nematode species (36,37)] a second SL exists, SL2, which is *trans*-spliced to 8.4–9.2% of mRNA speices, and which is used only for resolving polycistronic transcripts (38). In other organisms—flatworms (17), *Oikopleura* (10) and probably more-distantly-related nematodes (16)—one and the same SL is used both to *trans*-spliced monocistronic genes and for resolving polycistronic transcripts, and our results strongly indicate that *Ciona* also has a single major SL.

Global parameters of the *trans*-spliced and non-*trans*-spliced mRNA populations

The fact that SL-full-length ESTs comprised 27% (563/2078) of the total full-length EST population suggests that 27% of the total population of mRNA molecules are *trans*-spliced. However, this assumes that *trans*-spliced and non-*trans*-spliced mRNAs were amplified and cloned with similar efficiency by the oligo-capping procedure. We were able to independently confirm that this was indeed the case by additional analysis of existing EST data from unbiased conventional (and hence 5′-incomplete) cDNA libraries (6) representing the same developmental stages we had used to make the oligo-capping full-length EST library. Extensive linear overlap of conventional ESTs with our full-length EST sequences allowed us to identify and count, in the EST dataset, the numbers of molecules corresponding to the *trans*-spliced (SL-full-length EST) and non-*trans*-spliced (non-SL-full-length EST) mRNAs. The 668 genes in our full-length EST dataset were represented by a total of 19 242 cDNA molecules in the conventional EST libraries, and, of these, 5984 or 31%, correspond to *trans*-spliced mRNAs. The excellent agreement of this figure with the 27% estimated from direct analysis of the full-length EST library indicates that, compared with conventional cDNA cloning, the oligo-capping procedure did not introduce a strong bias for or against the *trans*-spliced mRNA subpopulation. This finding differs from that of a study of the cestode flatworm *Echinococcus*, which reported that the oligo-capping method was strongly biased against *trans*-spliced mRNAs (39). Presumably this bias is based on a feature of *trans*-spliced *Echinococcus* mRNAs that is not shared with *trans*-spliced *Ciona* mRNAs.

The gene populations represented by SL-full-length ESTs and non-SL-full-length ESTs were almost entirely distinct; only 14/668 genes (2.1%) were represented by both full-length EST types (Supplementary Table S2; Supplementary Figure S2). Even among those genes represented by 3 or more full-length ESTs, the vast majority (130/140 = 92%) were represented either entirely by SL-full-length ESTs or entirely by non-SL-full-length ESTs. This fact establishes several points. First, because they formed a distinct sequence set, the multiply-represented non-SL-full-length ESTs were not simply failed 5′-incomplete copies of *trans*-spliced mRNAs, but apparently represent a distinct population of bona fide non-*trans*-spliced mRNA molecules. Moreover, several lines of evidence indicated that at most a very small fraction of even singly-represented non-SL-full-length ESTs in our library could be 5′-incomplete copies of *trans*-spliced mRNAs: (i) a superabundant *trans*-spliced mRNA was represented by 123 SL-full-length ESTs and zero non-SL-full-length ESTs (Supplementary Table S3) and (ii) only a small minority (3/44) of non-SL-full-length ESTs derived from the 14 dual *trans*-spliced/non-*trans*-spliced genes had a structure that could be compatible with an artifactual origin as 5′-incomplete copies of the corresponding *trans*-spliced mRNA (Supplementary Figure S2E and F). A second point is that *trans*-splicing in *Ciona* is largely efficient; most of the mRNA molecules derived from *trans*-spliced genes are in fact *trans*-spliced (from genes represented by three or more full-length ESTs including at least one SL-full-length EST, we obtained a total of 256 full-length ESTs of which 215 were SL-full-length ESTs and 41 were non-SL-full-length ESTs, suggesting a *trans*-splicing efficiency for *trans*-spliced genes of at least 84%). A third point is that the *trans*-spliced and non-*trans*-spliced gene populations each represent approximately one-half of the total gene number: 332/668 full-length EST-represented genes were *trans*-spliced, including the small number of dual *trans*-spliced/non-*trans*-spliced genes in this category. This corresponds to 50%, with the 99% confidence interval being 45–55%.

The difference between the proportion of expressed genes that give rise to *trans*-spliced mRNAs (∼50%) and the proportion of accumulated mRNA molecules that are *trans*-spliced (27–31%, as estimated above) indicates that, on average, individual *trans*-spliced mRNA species are 2–3-fold less abundant than individual non-*trans*-spliced mRNA species. This unexpected difference was not due to the presence of a small number of unusually-abundant non-*trans*-spliced mRNAs, but appeared to reflect general population features (Supplementary Table S3).

The biological importance of SL *trans*-splicing is not entirely understood and it is sometimes discussed to be related with gene functions. As our study identified a significant number of both *trans*-spliced and non-*trans*-spliced genes within the same species, it provided a unique opportunity to assess whether these might represent distinct functional classes. We assessed gene function through Gene Ontology (GO) annotations (40). Only one out of 42 assigned GO terms in the third rank of the molecular function category showed a clear difference in representation in the two gene sets, suggesting extensive functional overlap of *trans*-spliced and non-*trans*-spliced genes. However, we noted a significant differential representation of ribosomal protein genes (GO: 0003735, structural constituent of ribosome), which were preferentially encoded by non-*trans*-spliced genes. Detailed inspection showed that seventy-six of the 79 ribosomal protein genes we could identify in the *Ciona* genome were represented in our full-length EST set (Supplementary Table S4). None were exclusively *trans*-spliced, 5 were dual *trans*-spliced/non-*trans*-spliced genes and the remaining 71 were non-*trans*-spliced genes. The biological significance of this marked preference of *Ciona* ribosomal protein genes for non-*trans*-spliced gene expression is unclear.

Polycistronic transcription units in *Ciona*

Our studies also revealed evidence for SL-resolved operons in the *Ciona* genome. In SL-resolved operons in other organisms, the member genes are transcribed in the same direction and are very close neighbours in the genome [intercistronic regions are most often ∼100 nt in the nematode *Caenorhabditis* (16), and 23–30 nt in *Oikopleura* (10)]. In addition, whereas the 5′-most cistron may be *trans*-spliced or not, all downstream cistrons are *trans*-spliced. We assessed the possible presence of similarly organized genes in *Ciona* by mapping our SL-full-length ESTs onto the genome and asking whether any of them were located very close to upstream gene models. For this analysis, we used an improved set of gene models recently predicted by the grailexp program (41) based on ∼680 000 ESTs and ∼6500 full-insert cDNA clone sequences (30). We were able to identify upstream gene models for 310 of the 332 SL-full-length EST-represented *trans*-spliced genes. In 173 of these gene pairs (group I) the genes were in the same transcriptional orientation and in 137 pairs they were in opposite orientation (group II). Thus operons would be expected to be found in group I but not in group II, nor in same-orientation (group III) or opposite-orientation (group IV) gene pairs made up of non-SL-full-length EST-represented genes and their upstream neighbours. Indeed, as shown in Figure 1, we found that very short intergenic regions (<100 nt) were moderately common in group I (9.2%, or 16 gene pairs, termed candidate operons 1–16) but were rare in groups II, III and IV (<1.1%), consistent with the expectation that some group I pairs might represent SL-resolved operons. As shown below (see Figure 3 and Supplementary Figure S3), detailed inspection demonstrated that these candidate operons lacked intercistronic regions, which strongly supports the hypothesis of polycistronic transcription.

We also carried out a full-length EST-independent whole-genome computational search for neighbouring gene-model pairs having operon-like properties, i.e. in the same orientation and separated by <100 nt. This yielded a total of 352 candidate operons, including operons 1–12 previously identified in the full-length EST-based analysis (Supplementary Table S5; operons 13–16 were missed in this screen because their downstream member genes were not accurately represented by the gene model set we used). Most candidate operons (328/352 = 93%) consisted of two genes, but some appeared to contain three (21 cases) or four (3 cases) genes (global average 2.08 genes per operon). Recognition of candidate operons depends on accuracy in the gene model predictions. Hence, the population of 352 candidate operons is likely to be incomplete, although our gene model accuracy estimates suggest that the missing fraction would be a minority.

To obtain direct evidence for polycistronic transcription of candidate operons we surveyed conventional EST data searching for cDNA clones representing incompletely processed, unresolved dicistronic transcripts, because such dicistronic ESTs that were experimentally obtained are theoretically the exact equivalent of the RT–PCR amplified dicistronic transcripts. Although, in *Caenorhabditis*, unresolved precursors are rare and in many cases undetectable (16), the great depth of available *Ciona* EST data raised the possibility that even rare RNA species might be found. In a grailexp model-based scan of the EST dataset, we found eight putative dicistronic transcripts (operons 33–40, Supplementary Table S6), all corresponding to candidate operons previously found in the whole-genome scan. In the course of other studies, we also found three additional putative dicistronic transcripts (operons 41–43, Supplementary Table S6), which had not been discovered in the grailexp model-based dicistronic EST search or whole-genome scan because one or both genes were not accurately represented by grailexp models.

In the case of operon 41 the evidence for SL-resolved polycistronic transcription is most extensive and is summarized in Figure 2. Operon 41 consists of two adjacent genes in the same transcriptional orientation. The downstream gene encodes a homologue of the GTP-binding nuclear protein Ran, while the upstream gene encodes a different protein similar to hypothetical proteins in other animals. The adjacent positioning of the genes is not a genome assembly artifact; the raw genome shotgun sequence data, which were obtained in the previous study (7), included eight separate reads across the intercistron boundary (3 of which are indicated in Figure 2). The accuracy of intron–exon structure of these

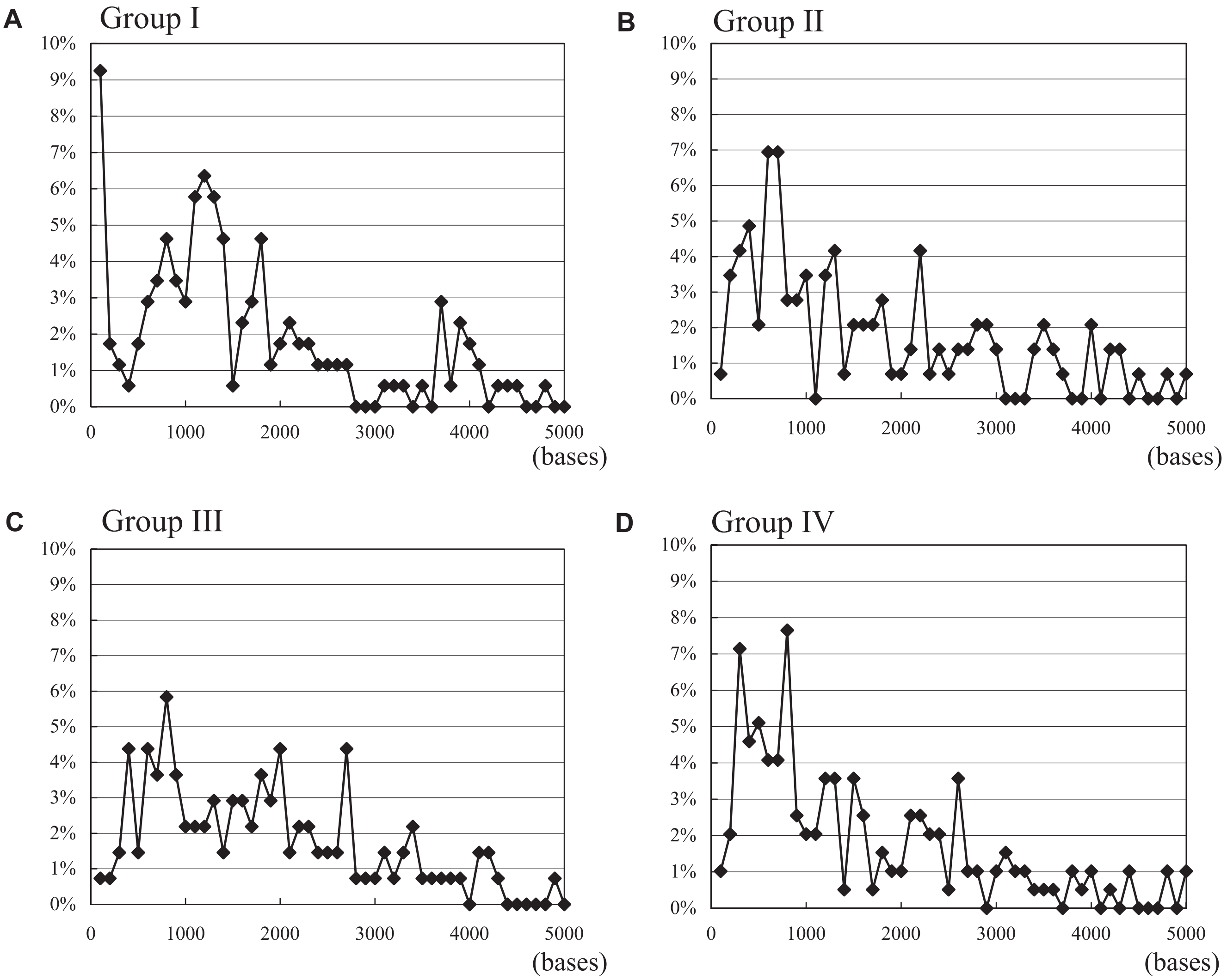


Figure 1. Distances between 5′ ends of genomically-mapped full-length ESTs and the nearest end of their 5′-neighbouring gene models, plotted in 100 base-windows. (**A**) SL-full-length ESTs whose 5′-neighbouring gene is transcribed in the same direction (group I), (**B**) SL-full-length ESTs whose 5′-neighbouring gene is transcribed in the opposite direction (group II), (**C**) non-SL-full-length ESTs whose 5′-neighbouring gene is transcribed in the same direction (group III) and (**D**) non-SL-full-length ESTs whose 5′-neighbouring gene is transcribed in the opposite direction (group IV). Note that only in (A) (group I) are gene pairs separated by <100 bases (first data point) well-represented. The number of gene pairs representing each distance interval is reported as a percentage of the total number of gene pairs in each group.

genes, and the fact that they are independent genes at the protein level, is attested by the existence in the EST dataset of hundreds of cDNA clones corresponding to separate monocistronic polyadenylated mature mRNAs (summarized in Figure 2). In addition to the monocistronic mRNAs, the EST data included two cDNA molecules, citb076c21 and cima003i16, that corresponds to an unresolved dicistronic transcript of operon 41, thereby providing direct experimental evidence for polycistronic transcription. In both cases the 3′-EST sequencing run encoded Ran, while the 5′-EST encoded the other protein. Finally, Ran mRNA was represented by three SL-full-length ESTs in our full-length EST analysis (Figure 2) and is therefore a *trans*-spliced mRNA species. Thus operon 41 has every feature expected of an SL-resolved dicistronic locus. Moreover, detailed features of the intercistron boundary in this and other putative incompletely-processed operons (see below) makes it very unlikely that the dicistronic transcripts could represent aberrant readthrough transcripts linking two transcriptionally independent genes (see Discussion).

Operon 43 also has all of the sorts of evidences available in the case of operon 41, while other individual operons may lack one or more of these pieces of evidence. However, the collective data strongly argue that candidate operons form a coherent set of genomic entities, most or all of which are SL-resolved polycistrons. We found dicistronic ESTs for nine candidate operons in addition to operons 41 and 43, and in each case the accuracy of the genome assembly was confirmed by multiple shotgun genome sequence reads, and multiple cDNAs representing the separate moncistronic mRNAs were found among the EST data. In these cases the downstream genes did not happen to be among the 4.2% of genes sampled in our full-length EST set, so we could not confirm that the downstream mRNAs were *trans*-spliced.

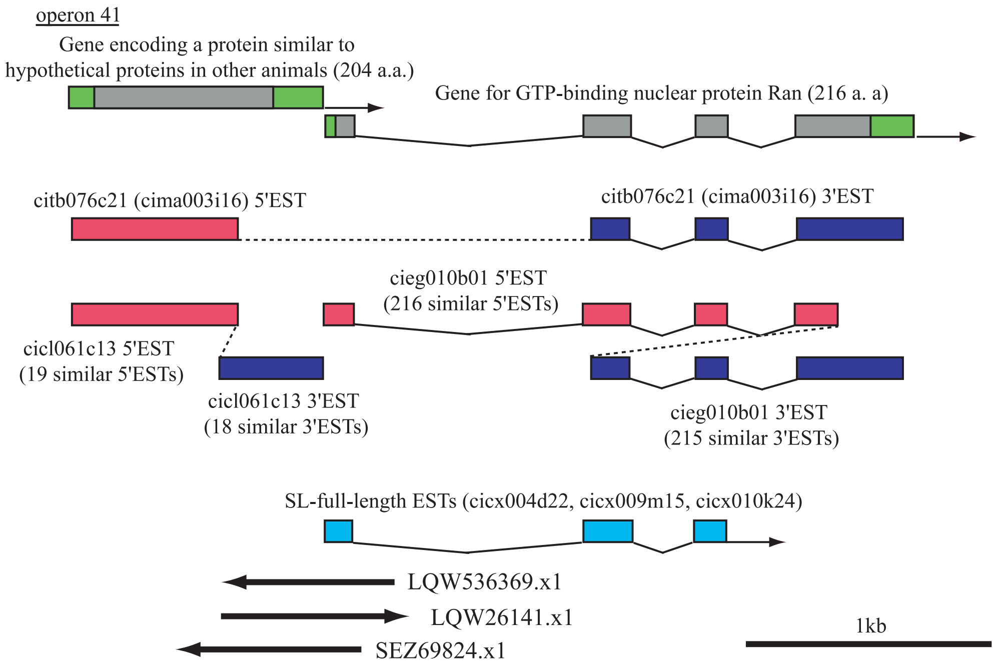

operon 41
Gene encoding a protein similar to hypothetical proteins in other animals (204 a.a.)
Gene for GTP-binding nuclear protein Ran (216 a. a)
citb076c21 (cima003i16) 5'EST
citb076c21 (cima003i16) 3'EST
cieg010b01 5'EST (216 similar 5'ESTs)
cicl061c13 5'EST (19 similar 5'ESTs)
cicl061c13 3'EST (18 similar 3'ESTs)
cieg010b01 3'EST (215 similar 3'ESTs)
SL-full-length ESTs (cicx004d22, cicx009m15, cicx010k24)
LQW536369.x1
LQW26141.x1
SEZ69824.x1
1kb

Figure 2. Dicistronic and moncistronic conventional ESTs, and SL-full-length ESTs, representing operon 41. The top part of the figure is a schematic depiction of the genomic DNA in terms of the intron/exon (lines/boxes) structures for two adjacent genes. The genes, coding for a protein similar to hypothetical proteins in other animals and a GTP-binding nuclear protein Ran, are immediately adjacent, so they are shown on separate lines for clarity. Protein-coding and non-coding regions are shown by grey and green. Below the genomic DNA depiction are diagrams showing conventional EST cDNA clones aligning with exons in this region. Each cDNA clone is represented by two EST sequencing runs, a 5′-EST (red) and a 3′ EST (blue), which are joined by dashed lines and which in some cases overlap. The first cDNA clone depicted, citb076c21, is a dicistronic transcript whose 5′-EST corresponds to the upstream gene and whose 3′ EST corresponds to the downstream gene (intron sequences were not present in the ESTs). An additional cDNA clone, cima003i16, was similar. Other cDNA clones depicted represent mature monocistronic mRNAs corresponding to either the upstream or downstream gene. The bottom depiction represents SL-full-length ESTs corresponding to the downstream gene, showing it to be *trans*-spliced. The genomic juxtaposition of these two cistrons is not due to an artifact of genome assembly, because eight raw whole-genome shotgun reads, of which three are depicted by black arrows at the bottom, can be aligned across the intercistron boundary.

However the following data showed that most, if not all, downstream genes in candidate operons are *trans*-spliced.

Our full-length EST dataset included 19 genes corresponding to downstream genes in candidate operons and in all 19 cases the full-length ESTs were SL-full-length ESTs (Table 2). The absence of non-SL-full-length ESTs is not informative in 7 of these cases because some operons were identified solely (operons 13–16) or perhaps partly (operons 41–43) on the basis of having a downstream *trans*-spliced gene. However the remaining 12 cases correspond to operons identified in the unbiased whole-genome gene model screen, and if there were no special relationship between candidate operons and *trans*-splicing, e.g. if these were simply pairs of transcriptionally independent genes that just happened to be unusually close together and in the same orientation, we would expect that the downstream genes would be *trans*-spliced or not in the proportions of the corresponding gene classes in the genome as a whole, i.e. half and half. Our finding that all (12/12) downstream genes are *trans*-spliced rules out the latter hypothesis (χ^2 test: $P \ll 0.01$), strongly supporting SL-resolved polycistronic transcription. Moreover, it indicates that erroneously-identified operons can be no more than a small minority of the candidate operon set. [As summarized in Table 2, the upstream genes in candidate operons included both *trans*-spliced (6 cases) and non-*trans*-spliced (14 cases) genes, which are consistent with all hypotheses, including SL-resolved operon gene expression].

Intercistron boundaries in operons

Detailed analysis of intercistron boundaries gave further evidence consistent with SL-resolved polycistronic transcription, and also indicated an unusual aspect of operon expression in *Ciona* as compared with other organisms. The downstream genes of operons 1–16 and 41–43 were represented by SL-full-length ESTs, which precisely localized the *trans*-splice acceptor sites. In each of these 19 cases, the 3′ end of the mRNA derived from the upstream gene could also be precisely localized by determining poly(A)-adjacent sequences in oligo(dT)-primed cDNA clones (see Materials and Methods). This precise localization of 3′ ends and *trans*-splice acceptor sites revealed that in each of these operons the upstream and downstream cistrons were directly juxtaposed, with no intercistronic DNA. [An apparent exception was operon 42 in which the downstream gene was represented by a single SL-full-length EST in which the SL was linked to exon 2; however, a gene-specific 5′-RACE experiment confirmed that in other mRNA molecules derived from the downstream gene, *trans*-splicing did occur precisely at the intercistron boundary marked by the upstream mRNA's 3′ end (data not shown)]. Two examples are shown in Figure 3 and the remaining 16 cases are shown in Supplementary Figure S3. In general, the first nucleotide of the downstream cistron, to which the SL was linked, was immediately adjacent to the last nucleotide of the upstream cistron, to which poly(A) was linked, so that the G residue of the

Table 2. Full-length EST representation of upstream and downstream genes in candidate operons

	Number of full-length ESTs			
	Upstream gene		Downstream gene	
Operon ID[a]	SL-full-length EST	Non-SL-full-length EST	SL-full-length EST	Non-SL-full-length EST
1	0	0	1	0
2	0	0	1	0
3	0	0	1	0
4	0	0	1	0
5	0	0	1	0
6	0	0	5	0
7	0	0	2	0
8	0	2	1	0
9	0	1	1	0
10	0	0	1	0
11	0	0	1	0
12	1	0	1	0
13	0	0	1	0
14	0	0	1	0
15	0	0	1	0
16	0	0	1	0
17	0	1	0	0
18	0	1	0	0
19	0	1	0	0
20	0	1	0	0
21	0	1	0	0
22	0	6	0	0
23	0	1	0	0
24	0	1	0	0
25	0	11	0	0
26	0	1	0	0
27	0	5	0	0
28	1	5	0	0
29	1	0	0	0
30	2	0	0	0
31	1	0	0	0
32	1	0	0	0
33	0	0	0	0
34	0	0	0	0
35	0	0	0	0
36	0	0	0	0
37	0	0	0	0
38	0	0	0	0
39	0	0	0	0
40	0	0	0	0
41	0	0	3	0
42	0	0	1	0
43	0	0	1	0
Total	7	38	26	0

[a] ID numbers assigned to candidate operons described in detail in the text. Genes within each operon are listed in Supplementary Table S5, except for operons 13–16 and 41–43. The downstream genes in candidate operons 13–16 are not accurately represented by gene models but were represented by cicx006d13 (operon13), cicx007d17 (operon14), cicx008h05 (operon15) and cicx009d07 (operon16). Operons 41–43 were found in other unrelated studies and their detailed features are shown in Supplementary Table S6.

AG dinucleotide immediately upstream of the *trans*-splice acceptor site served as the residue to which poly(A) was added on the upstream mRNA. This relationship applied to all operons shown in Figure 3 and Supplementary Figure S3, and was reflected in the majority (41/55) of upstream gene cDNAs. In eight operons, there was microheterogeneity of upstream cDNA 3′ ends; in some molecules, poly(A) addition had occurred at alternative sites 2–6 nt upstream—exceptionally in operon 15 as far as 67 nt. This 3′ end microheterogeneity could reflect partial nucleolytic processing of the upstream mRNA 3′ end prior to poly(A) addition, or a 2 nt mispriming shift by oligo(dT) on the . . . AGAAAAAAAA . . . . . . sequence during the first strand synthesis (‘AG’ corresponds to the acceptor site for the SL *trans*-splicing of the downstream cistron and the following ‘AAAAAAAA. . .. . .’ is the poly(A) tail of the upstream gene transcript).

In no other *trans*-splicing organism are the majority, or even a significant fraction, of operons known to lack intercistronic DNA. However it is interesting to note that a minor class of *Caenorhabditis* operons lack intercistronic sequences and are resolved by *trans*-splicing with SL1 rather than SL2 (42,43) (see Discussion).

DISCUSSION

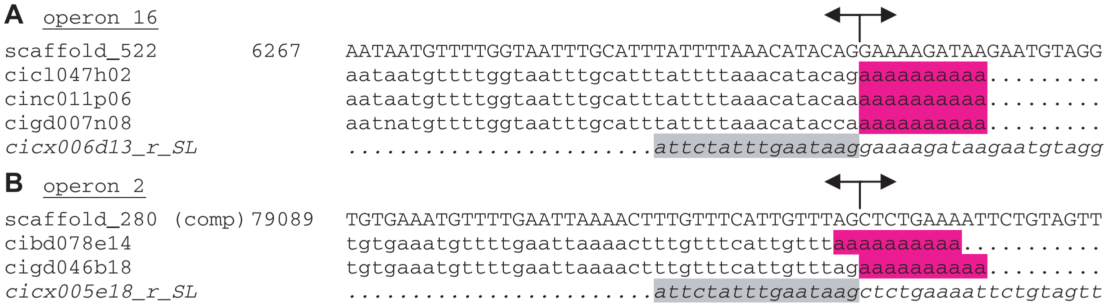
Our study showed that the genome of *Ciona* is composed of two nearly equal gene subsets with little overlap, one of which undergoes efficient pre-mRNA *trans*-splicing with the single major SL, while the other undergoes conventional non-*trans*-splicing expression. The ∼50% *trans*-spliced gene fraction we estimate for *Ciona* is lower than the 70–90% estimate for nematodes (25,44), but higher than the 12–24% estimate for *Oikopleura* (10). The significance of lineage-specific differences in overall *trans*-splicing levels is not clear, although it is likely that such differences could reflect, and/or contribute to, lineage-specific features of genome evolution.

The marked preferential encoding of ribosomal proteins by non-*trans*-spliced mRNAs in *Ciona* is the first correlation to be established in any organism between *trans*-splicing and gene functional classification. It is of interest that a different relationship appears to exist in *Oikopleura* where at least 43 ribosomal protein genes are *trans*-spliced (10). Ribosomal protein gene organization strategies also differ; at least 14 *Oikopleura* ribosomal protein genes are associated with candidate operons (10), but we found only 3 of the 79 identified *Ciona* ribosomal protein genes in candidate operons (data not shown). In *Caenorhabditis*, the *trans*-splicing status for most of the 115 known ribosomal protein genes has not been established, although a high proportion, ∼40%, have been shown to be associated with operons (16).

An unexpected finding of our study was that individual *trans*-spliced mRNA species in *Ciona* are, on average, 2–3-fold less abundant than non-*trans*-spliced mRNAs. This issue has not yet been investigated in any other organism and it will be of interest for future studies to establish if a similar abundance relationship holds in other species. A related question to be investigated is whether the mRNA abundance difference observed in *Ciona* is compensated by increased translational efficiency of *trans*-spliced mRNAs. (Differing studies in nematodes have reported that the translational efficiency of *trans*-spliced mRNAs is higher than (25) or similar to (26) that of non-*trans*-spliced mRNAs.)

Polycistronic transcription units

Our study led to the recognition of a set of 352 operons (mostly dicistronic) that are resolved by *trans*-splicing with the same SL used for monocistronic pre-mRNAs in the



```
A  operon 16
scaffold_522         6267   AATAATGTTTTGGTAATTTGCATTTATTTTAAACATACAGGAAAAGATAAGAATGTAGG
cicl047h02                  aataatgttttggtaatttgcatttattttaaacatacagaaaaaaaaaa.........
cinc011p06                  aataatgttttggtaatttgcatttattttaaacatacaaaaaaaaaaaa.........
cigd007n08                  aatnatgttttggtaatttgcatttattttaaacataccaaaaaaaaaaa.........
cicx006d13_r_SL             ........................attctatttgaataaggaaaagataagaatgtagg
B  operon 2
scaffold_280 (comp) 79089  TGTGAAATGTTTTGAATTAAAACTTTGTTTCATTGTTTAGCTCTGAAAATTCTGTAGTT
cibd078e14                  tgtgaaatgttttgaattaaaactttgtttcattgttttaaaaaaaaaa..........
cigd046b18                  tgtgaaatgttttgaattaaaactttgtttcattgtttagaaaaaaaaaa.........
cicx005e18_r_SL             ........................attctatttgaataagctctgaaaattctgtagtt
```

Figure 3. Cistrons in *Ciona* candidate operons are immediately juxtaposed with no intervening DNA. Genomic sequences (assembly version 1.0) and coordinates are shown on the top line of each panel, with scaffold name and nucleotide position indicated. ESTs mapping to this site are shown below; conventional ESTs for the upstream genes (lower case roman letters, poly(**A**)—10 residues shown—shaded in red) and SL-full-length ESTs for the downstream genes (lower case italic letters, SL sequence shaded in grey). The intercistron boundaries are indicated by arrows. In addition to the two examples shown here, 16 similarly-organized operons are shown in Supplementary Figure S3.

Ciona genome. In the best characterized cases, operons 41 and 43, the evidence for SL-resolved polycistronic transcription includes: (i) cDNA clones representing the unresolved dicistronic transcript, (ii) SL-full-length ESTs identifying the downstream member gene as being *trans*-spliced and (iii) an unusual genomic structure featuring the complete absence of intercistronic DNA. Because of the complete absence of intercistronic DNA, it is unlikely that the dicistronic transcripts could be aberrant read through transcripts of independent genes. Production of the 3′ end of the upstream mRNA by primary transcript cleavage, or by *trans*-splicing (see below), requires transcription across the intercistron boundary to generate the cleavage site. Likewise, production of the 5′ end of the mature *trans*-spliced downstream mRNA requires transcription across the intercistron boundary to generate the *trans*-splice acceptor site. Thus transcription across the intercistron boundary is not aberrant for an intercistronless operon, but is essential for the production of any mature monocistronic mRNA from the locus.

We found 19 candidate operons in which the downstream genes were represented by full-length ESTs (SL-full-length ESTs in all case). Given that the full-length EST set represents a 4.2% sampling of the entire genome, it can be estimated that the genome should contain a total of ~100/4.2 × 19 = 452 such intercistronless gene pairs. This number is higher than the 352 candidate operons identified in the whole genome gene-model scan, but the latter number was expected to be an underestimate because it made no allowance for inaccurate gene models. These findings together form a compelling case for the existence of a coherent set of ~350–450 SL-resolved (mostly dicistronic) operons in the *Ciona* genome.

Operon structure and resolution

Precise juxtaposition of upstream and downstream cistrons in *Ciona* operons suggests the possibility that the *trans*-splicing reaction itself could generate the 3′ end of the upstream mRNA. Indeed it seems very unlikely that an independent transcript cleavage mechanism could, in many independent cases, target precisely the same phosphodiester bond that is cleaved in the *trans*-splicing reaction. However, because the *trans*-splice acceptor site branch-point must necessarily reside within the upstream mRNA, the latter would presumably be released as a branched nucleic acid structure. Unless it was rapidly debranched by an unknown mechanism, this structure could have negative implications for mRNA function. Moreover, it seems certain that prior formation of the upstream mRNA's 3′ end (by a distinct mechanism) would preclude subsequent *trans*-splicing of the downstream cistron because the *trans*-splice branch-point, and acceptor site AG dinucleotide, will have been lost with the upstream mRNA. Thus expression of genes in operons that lack intercistronic sequences may be mutually exclusive in the sense that any given transcript molecule may be capable of producing either an upstream or a downstream mRNA, but not both [see also Ref. (42)]. As indicated by mRNA accumulation measured by EST counts (45), we found that operon member genes appeared to be independently expressed in overlapping patterns that were neither tightly coordinated nor markedly mutually exclusive, at least at the macroscopic level (data not shown). Further biochemical studies will be required to establish the mechanism of operon resolution in *Ciona* and the mechanisms that regulate the differential accumulation of mRNAs encoded by operons.

An observation that may be relevant to the mechanism of operon resolution is that most *Ciona* operons (16/18 cases shown in Figure 3 and Supplementary Figure S3) lack the canonical polyadenylation signal AAT(U)AAA within 40 nt upstream of the upstream cistron polyadenylation site (although this signal is used in a substantial fraction of *Ciona* genes (Y. Satou, unpublished data). *Oikopleura* operon upstream cistrons also lack the AATAAA signal (10). It can be noted that, whereas it might be desirable to have 100% efficient 3′ end formation in monocistronic genes, this is not the case for *Ciona* operon upstream cistrons because this would almost certainly preclude *trans*-splicing and expression of the downstream gene (see above). It might then be reasonable to expect mechanistic differences in 3′ end formation in monocistronic and operon gene classes.

Our findings add *Ciona* to the number of organisms, including *Oikopleura*, in which the same SL is used both for monocistronic and SL-resolved polycistronic expression. It seems increasingly likely that the occurrence in Clade V nematodes, such as *Caenorhabditis*, of an additional SL, SL2, specifically devoted to operon resolution is a highly specialized, lineage-specific feature (16). One of the functions of the intercistronic regions in *Caenorhabditis* operons is the specific recruitment of SL2 to downstream cistrons, in preference to the more abundant SL1 (23). However it is of great interest that a small minority of operons (at least 25 operons) in *Caenorhabditis* have little or no intercistronic sequence and are resolved by SL1, not SL2 (42, 43; T. Blumenthal, personal communication), similar to *Ciona* operons. *Ciona* is the first organism known in which the majority of operons entirely lack intercistronic DNA. Whether the intercistron-less operons of *Caenorhabditis* and *Ciona* represent convergent evolution or a common ancestral character is presently unknown, but is clearly a very important question. The presence of the canonical AATAAA cleavage/polyadenylation signal in upstream cistrons in the three fully characterized SL1-type operons in *Caenorhabditis* (42) suggests a possible mechanistic difference from the majority of *Ciona* operons. Further studies of the mechanisms of operon resolution in *Ciona* and in SL1-type operons of *Caenorhabditis*, and additional studies of operon structure and resolution in nematodes that lack SL2, would be of great interest in this regard.

SL *trans*-splicing evolution in chordates

The presence of SL *trans*-splicing and its use to resolve polycistronic transcripts in both *Ciona* and distantly-related *Oikopleura* suggest that these were ancestral tunicate features. However this does not establish whether *trans*-splicing and operons arose early within the tunicate lineage following the divergence of the vertebrate lineage, or were present in the ancestral chordate and were secondarily lost in the vertebrate lineage. It is interesting that despite overall similarities, there are numerous differences between SL *trans*-splicing in *Ciona* and *Oikopleura*. These include the length and sequence of the SL itself (9,10), the proportion of genes that are *trans*-spliced, the presence in *Oikopleura*, but not in *Ciona*, of 23–30 nt intercistronic sequences in candidate operons, and the preferential encoding of ribosomal proteins by *trans*-spliced and/or operon associated genes (*Oikopleura*) versus non-*trans*-spliced and non-operon-associated genes (*Ciona*). These differences, and other evolutionary genomic differences, such as the small genome size and short intron lengths in *Oikopleura* (46) suggest that further comparative studies of tunicates may be particularly informative about evolutionary aspects of *trans*-splicing and its relationship to genome evolution.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

The authors appreciate the technical assistance of Mr Daisuke Miyamura and Dr Takeshi Kawashima. This research was supported by Grants-in-Aid from the Ministry of Education, Science, Sports, Culture and Technology, Japan to Y.S. (17071020 and 17687022). This research was also supported in part by CREST project to N.S. K.E.M.H. was supported by a grant from NSERC. Funding to pay the Open Access publication charges for this article was provided by the Ministry of Education, Science, Sports, Culture and Technology, Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Aparicio,S., Chapman,J., Stupka,E., Putnam,N., Chia,J.M., Dehal,P., Christoffels,A., Rash,S., Hoon,S., Smit,A. *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, **297**, 1301–1310.
2. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
3. International human genome sequencing consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
4. Satoh,N. (2003) The ascidian tadpole larva: comparative molecular development and genomics. *Nature Rev. Genet.*, **4**, 285–295.
5. Satoh,N., Satou,Y., Davidson,B. and Levine,M. (2003) *Ciona intestinalis*: an emerging model for whole-genome analyses. *Trends Genet.*, **19**, 376–381.
6. Satou,Y., Yamada,L., Mochizuki,Y., Takatori,N., Kawashima,T., Sasaki,A., Hamaguchi,M., Awazu,S., Yagi,K., Sasakura,Y. *et al.* (2002) A cDNA resource from the basal chordate *Ciona intestinalis*. *Genesis*, **33**, 153–154.
7. Dehal,P., Satou,Y., Campbell,R.K., Chapman,J., Degnan,B., De Tomaso,A., Davidson,B., Di Gregorio,A., Gelpke,M., Goodstein,D.M. *et al.* (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science*, **298**, 2157–2167.
8. Nilsen,T.W. (1993) *Trans*-splicing of nematode premessenger RNA. *Annu. Rev. Microbiol.*, **47**, 413–440.
9. Vandenberghe,A.E., Meedel,T.H. and Hastings,K.E. (2001) mRNA 5′-leader *trans*-splicing in the chordates. *Genes Dev.*, **15**, 294–303.
10. 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.
11. Nilsen,T.W. (2001) Evolutionary origin of SL-addition *trans*-splicing: still an enigma. *Trends Genet.*, **17**, 678–680.
12. 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.
13. Hastings,K.E. (2005) SL *trans*-splicing: easy come or easy go? *Trends Genet.*, **21**, 240–247.
14. Pouchkina-Stantcheva,N.N. and Tunnacliffe,A. (2005) Spliced leader RNA mediated *trans*-splicing in phylum Rotifera. *Mol. Biol. Evol.*, **22**, 1482–1489.
15. Spieth,J., Brooke,G., Kuersten,S., Lea,K. and Blumenthal,T. (1993) Operons in *C.elegans*: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell*, **73**, 521–532.
16. Blumenthal,T. and Gleason,K.S. (2003) *Caenorhabditis elegans* operons: form and function. *Nature Rev. Genet.*, **4**, 112–120.
17. 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.
18. Johnson,P.J., Kooter,J.M. and Borst,P. (1987) Inactivation of transcription by UV irradiation of *T. brucei* provides evidence for a multicistronic transcription unit including a VSG gene. *Cell*, **51**, 273–281.
19. Muhich,M.L. and Boothroyd,J.C. (1988) Polycistronic transcripts in trypanosomes and their accumulation during heat shock: evidence for a precursor role in mRNA synthesis. *Mol. Cell. Biol.*, **8**, 3837–3846.
20. Tschudi,C. and Ullu,E. (1988) Polygene transcripts are precursors to calmodulin mRNAs in trypanosomes. *EMBO J.*, **7**, 455–463.
21. Vanhamme,L. and Pays,E. (1995) Control of gene expression in trypanosomes. *Microbiol Rev.*, **59**, 223–240.

22. Campbell,D.A., Thomas,S. and Sturm,N.R. (2003) Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect*., **5**, 1231–1240.
23. Huang,T., Kuersten,S., Deshpande,A.M., Spieth,J., MacMorris,M. and Blumenthal,T. (2001) Intercistronic region required for polycistronic pre-mRNA processing in *Caenorhabditis elegans*. *Mol. Cell. Biol*., **21**, 1111–1120.
24. Lercher,M.J., Blumenthal,T. and Hurst,L.D. (2003) Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. *Genome Res*., **13**, 238–243.
25. 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.
26. 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.
27. Blumenthal,T. (1995) *Trans*-splicing and polycistronic transcription in *Caenorhabditis elegans*. *Trends Genet*., **11**, 132–136.
28. Chomczynski,P. and Sacchi,N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem*., **162**, 156–159.
29. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred I. Accuracy assessment. *Genome Res*., **8**, 175–185.
30. Satou,Y., Kawashima,T., Shoguchi,E., Nakayama,A. and Satoh,N. (2005) An integrated database of the ascidian, *Ciona intestinalis*: towards functional genomics. *Zoolog. Sci*., **22**, 837–843.
31. Kent,W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
32. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
33. Satou,Y., Takatori,N., Yamada,L., Mochizuki,Y., Hamaguchi,M., Ishikawa,H., Chiba,S., Imai,K., Kano,S., Murakami,S.D. *et al*. (2001) Gene expression profiles in *Ciona intestinalis* tailbud embryos. *Development*, **128**, 2893–2904.
34. D'Alessio,J.M. and Gerard,G.F. (1988) Second-strand cDNA synthesis with *E. coli* DNA polymerase I and RNase H: the fate of information at the mRNA 5′ terminus and the effect of E. coli DNA ligase. *Nucleic Acids Res*., **16**, 1999–2014.
35. Suzuki,Y., Taira,H., Tsunoda,T., Mizushima-Sugano,J., Sese,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Morishita. *et al*. (2001) Diverse transcriptional initiation revealed by fine large-scale mapping of mRNA start sites. *EMBO Rep*., **2**, 388–393.
36. Blaxter,M.L., De Ley,P., Garey,J.R., Liu,L.X., Scheldeman,P., Vierstraete,A., Vanfleteren,J.R., Mackey,L.Y., Dorris,M., Frisse,L.M. *et al*. (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature*, **392**, 71–75.
37. Lee,K.Z. and Sommer,R.J. (2003) Operon structure and *trans*-splicing in the nematode *Pristionchus pacificus*. *Mol. Biol. Evol*., **20**, 2097–2103.
38. Blumenthal,T., Evans,D., Link,C.D., Guffanti,A., Lawson,D., Thierry-Mieg,J., Thierry-Mieg,D., Chiu,W.L., Duke,K., Kiraly,M. *et al*. (2002) A global analysis of *Caenorhabditis elegans* operons. *Nature*, **417**, 851–854.
39. Fernandez,C., Gregory,W.F., Loke,P. and Maizels,R.M. (2002) Full-length-enriched cDNA libraries from *Echinococcus granulosus* contain separate populations of oligo-capped and *trans*-spliced transcripts and a high level of predicted signal peptide sequences. *Mol. Biochem. Parasitol*., **122**, 171–180.
40. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.
41. Uberbacher,E.C., Xu,Y. and Mural,R.J. (1996) Discovering and understanding genes in human DNA sequence using GRAIL. *Meth. Enzymol*., **266**, 259–281.
42. Williams,C., Xu,L. and Blumenthal,T. (1999) SL1 *trans* splicing and 3′-end formation in a novel class of *Caenorhabditis elegans* operon. *Mol. Cell. Biol*., **19**, 376–383.
43. Hengartner,M.O. and Horvitz,H.R. (1994) *C. elegans* cell survival gene ced-9 encodes a functional homolog of the mammalian proto-oncogene bcl-2. *Cell*, **76**, 665–676.
44. Zorio,D.A., Cheng,N.N., Blumenthal,T. and Spieth,J. (1994) Operons as a common form of chromosomal organization in *C. elegans*. *Nature*, **372**, 270–272.
45. Satou,Y., Kawashima,T., Kohara,Y. and Satoh,N. (2003) Large scale EST analyses in *Ciona intestinalis*: its application as Northern blot analyses. *Dev. Genes. Evol*., **213**, 314–318.
46. Seo,H.C., Kube,M., Edvardsen,R.B., Jensen,M.F., Beck,A., Spriet,E., Gorsky,G., Thompson,E.M., Lehrach,H., Reinhardt,R. *et al*. (2001) Miniature genome in the marine chordate *Oikopleura dioica*. *Science*, **294**, 2506.