Identification and analysis of internal promoters in *Caenorhabditis elegans* operons

Peiming Huang, Erin D. Pleasance, Jason S. Maydan, et al.

Genome Res. published online August 21, 2007
Access the most recent version at doi:10.1101/gr.6824707

Letter

Identification and analysis of internal promoters in *Caenorhabditis elegans* operons

Peiming Huang,[1] Erin D. Pleasance,[1] Jason S. Maydan,[2] Rebecca Hunt-Newbury,[2] Nigel J. O'Neil,[3] Allan Mah,[4] David L. Baillie,[4] Marco A. Marra,[1,3] Donald G. Moerman,[2] and Steven J.M. Jones[1,3,5]

[1]*Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia V5Z 1L3, Canada;* [2]*Department of Zoology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada;* [3]*Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada;* [4]*Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada*

The current *Caenorhabditis elegans* genomic annotation has many genes organized in operons. Using directionally stitched promoter::GFP methodology, we have conducted the largest survey to date on the regulatory regions of annotated *C. elegans* operons and identified 65, over 25% of those studied, with internal promoters. We have termed these operons "hybrid operons." GFP expression patterns driven from internal promoters differ in tissue specificity from expression of operon promoters, and serial analysis of gene expression data reveals that there is a lack of expression correlation between genes in many hybrid operons. The average length of intergenic regions with putative promoter activity in hybrid operons is larger than previous estimates for operons as a whole. Genes with internal promoters are more commonly involved in gene duplications and have a significantly lower incidence of alternative splicing than genes without internal promoters, although we have observed almost all *trans*-splicing patterns in these two distinct groups. Finally, internal promoter constructs are able to rescue lethal knockout phenotypes, demonstrating their necessity in gene regulation and survival. Our work suggests that hybrid operons are common in the *C. elegans* genome and that internal promoters influence not only gene organization and expression but also operon evolution.

[Supplemental material is available online at www.genome.org.]

Genes that participate in a particular biological process in bacteria and archaea are often organized in an operon for coordinated regulation of gene expression (Jacob and Monod 1961; Langer et al. 1995). Genes that are cotranscribed as polycistronic messages have also been observed in many eukaryotic genomes such as flies (Andrews et al. 1996; Liu et al. 2000), mammals (Sloan et al. 1999; Corcoran et al. 2004), plants (Garcia-Rios et al. 1997; Thimmapuram et al. 2005), trypanosomes (Imboden et al. 1987), and nematodes (Spieth et al. 1993). In comparison to other eukaryotes, *Caenorhabditis elegans*, *Caenorhabditis briggsae*, and other nematodes have a relatively large number of genes grouped in operons (Evans and Blumenthal 2000). In *C. elegans*, at least 15% of genes are believed to be involved in polycistronic units, and a genomic comparison between *C. elegans* operons and those of *C. briggsae* has indicated that >90% of them are conserved (Blumenthal et al. 2002; Blumenthal and Gleason 2003). As in prokaryotes, eukaryotic genes packaged in a polycistronic unit have the ability to be transcribed from a single promoter at the 5′-end, producing a single primary pre-mRNA. Some eukaryotes such as flies and mammals, however, are able to reinitiate translation of downstream genes through internal ribosome entry sites (Macejak and Sarnow 1991; Ye et al. 1997), whereas *C. elegans* uses *trans*-splice machinery to separate a polycistronic pre-mRNA into monocistronic mRNAs rather than translate them as a single unit (Spieth et al. 1993; Blumenthal et al. 2002). Nearly 70% of *C. elegans* pre-mRNAs are *trans*-spliced, and ~25% of these pre-mRNAs are predicted to be within operons (Zorio et al. 1994).

The *trans*-splicing of a small RNA molecule, the spliced leader (SL), to the 5′-end of a pre-mRNA was first discovered in trypanosomes (Murphy et al. 1986). Although it has been subsequently observed in nematodes, flatworms, and some other phyla since then, the evolutionary origin of SL *trans*-splicing is still unknown (Nilsen 2001). The mechanism of *trans*-splicing, however, closely resembles that of *cis*-splicing (Blumenthal 2005; MacMorris et al. 2007; Saldi et al. 2007). In *C. elegans* there are two major types of spliced leader, SL1 and SL2. Both are 22 nt long and possess a trimethyl guanosine cap, which is used to protect 5′-ends of mature mRNAs and facilitate the binding of the translational apparatus. SL1 is the major form used in *trans*-splicing for the 5′-ends of pre-mRNAs. It has been proposed that SL2 is exclusively involved in the *trans*-splicing of downstream genes (non-leading, or distal genes) within operons, while some operon downstream genes are also *trans*-spliced to SL1 (Spieth et al. 1993; Zorio et al. 1994). The discovery of SL2 *trans*-splicing helped to unravel the operon phenomenon in *C. elegans* (Huang and Hirsh 1989; Spieth et al. 1993). Therefore, locating SL2 sites among closely spaced genes with the same orientation has become the main criterion for operon annotation in *C. elegans*. The average length of intergenic spaces between adjacent genes in *C. elegans* polycistronic units is ~100 bp, with some as high as 300–400 bp. The significance of the variation in gene spacing of annotated operons is not known (Zorio et al. 1994).

It has been established as a general principle that gene duplication plays an important role in the evolution of genetic

[5]Corresponding author.
E-mail sjones@bcgsc.ca; fax (604) 876-3561.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6824707.

variation and new protein functionalities (Ohno 1970). In *C. elegans*, most gene duplication events involve only one or two genes (Semple and Wolfe 1999). Duplicated genes can be expected to retain their functionality as long as the duplications retain the transcriptional ability of the corresponding parent copies. The successful duplication of operon downstream genes, however, is limited by the operon structure because duplicates need to contain a sufficient number of upstream elements, including operon regulatory regions, in order to be transcribed (Lercher et al. 2003) or move into an actively transcribed region to contribute to a phenotype.

Through the analysis of microarray expression data, Lercher et al. (2003) observed that genes grouped in operons showed a higher degree of correlation in their expression than that observed between non-operon genes. The expression correlation was not very strong, however, and it weakened dramatically with increasing intergenic distances in operons. Lercher et al. (2003) concluded that coexpression of operon genes was affected by some "distance-dependent regulation." One contributing factor, which influences gene expression analysis, is that transcripts undergo differential rates of decay (Blumenthal 2004). Gene expression analysis in other eukaryotes, such as mouse and human, has shown that genes with similar expression profiles tend to be physically close together despite the lack of operon structures within these genomes (Singer et al. 2005), an effect possibly mediated by other genomic features such as bidirectional promoters (Trinklein et al. 2004).

Green fluorescent protein (GFP) reporter gene methodology was first introduced in *C. elegans* in 1994 (Chalfie et al. 1994). Since then, it has become an important monitoring tool in biological research, used in the study of in vivo subcellular protein localization and protein dynamics, as well as in the analysis of spatial and temporal gene expression in living cells and tissues. Promoter::GFP fusions have been widely used for the identification and characterization of promoters in both prokaryotes and eukaryotes (Lu et al. 2004; Zhao et al. 2004). Although this reporter system may not precisely record the temporal and spatial patterning of the gene expression because of the exclusion of more distal regulatory elements from the reporter construct, it does provide evidence that a segment of DNA under study has the capability to autonomously promote transcription in vivo. An enhancer cannot function to promote transcription without the presence of basal promoter elements. Therefore, we have used this system in this study for the identification of internal promoters within *C. elegans* operons.

Operons with internal promoters for downstream genes, which we have termed "hybrid operons," have been identified and characterized in many bacterial genomes (Horowitz and Platt 1983). Internal promoters have also been suggested in *C. elegans* operons (Spieth et al. 1993; Zhao et al. 2004). In this study, we present our analysis of *C. elegans* operons based on WormBase annotations (WS135) and our gene-specific directionally stitched promoter::GFP fusion results. We suggest that hybrid operons are common in the *C. elegans* genome.

Results

Identification of putative internal promoters in hybrid operons

We discovered internal promoters by searching for promoter activity located in intergenic regions of WormBase annotated operons. From a set of 2448 directionally stitched promoter::GFP constructs, we have identified 979 putative gene-specific promoters able to drive reporter GFP expression (McKay et al. 2003). Within this set, we identified internal promoters for a total of 66 downstream genes derived from 65 hybrid operons. These are summarized in Supplemental Table 1. Of the remaining GFP constructs that did not demonstrate promoter activity by GFP expression, 172 were at operon downstream locations and therefore have no evidence for containing internal promoters. Overall, 27.7% of the 238 genes annotated as being downstream in operons and tested in this analysis show the ability to promote transcription independently. This compares to an overall success rate for genes not predicted to be in operons of 41.3% (913 out of 2210).

As an example, Figure 1A shows details of the *klp-8* operon (CEOPX040). We have constructed promoter::GFP fusions for both the leading gene C15C7.2 and the downstream gene C15C7.1 of the *klp-8* operon, and each is able to drive expression of the reporter gene in different tissues (Fig. 1B), indicating the presence of an internal promoter. The same is true for the CEOP3332 operon (Fig. 1C,D). The expression patterns driven by other promoter pairs for which we have data on both leading genes and downstream genes in the same operons are shown in Supplemental Figure 1.

To further establish that internal promoters in our study are active in vivo outside the context of a reporter assay, we also looked for examples in which the downstream gene could rescue a lethal mutant phenotype in the absence of promoter sequence from the upstream gene. It has been reported multiple times in the literature that *unc-37* mutants can be rescued without the promoter of the upstream gene (Pflugrad et al. 1997; Chang et al. 2003). Also, we were able to rescue two independently derived alleles of the homozygous lethal *let-721* mutants with only 1137 bp upstream of the translational initiation start site, which excludes the upstream promoter.

Finally, to confirm the observed negative GFP expression, a deletion analysis was conducted in order to capture the largest possible upstream regulatory sequences for downstream genes. For 12 operons selected in the test set described in Methods, we designed constructs that include the entire upstream genes plus the entire intergenic regions, but exclude the promoters for the upstream genes, which were deleted. The control set contains the sequences from the test set plus the promoter sequences from the leading genes. Our results showed that the constructs from the control set still drove positive GFP expression, while constructs from 11 out of the 12 downstream genes in the test set were unable to drive GFP expression, indicating that the majority of the negative downstream genes in our original experiment lack internal promoters.

Properties of putative internal promoters in hybrid operons

Based on WormBase annotations (Stein et al. 2001), out of 66 downstream genes with internal promoters studied (Table 1), 38% use a mixture of SL1/SL2 and 56% use SL2 only for *trans*-splicing. The remaining 6%, however, do not yet have evidence for the presence of spliced leaders. Our own analysis of SL1 and SL2 *trans*-splicing (P. Huang, unpubl.), based on the method described by Hwang et al. (2004), indicates that there are an additional nine downstream genes that are actually *trans*-spliced to both SL1 and SL2, bringing the total to 34, or 51.5%.

Of 172 operon downstream genes without internal promot-

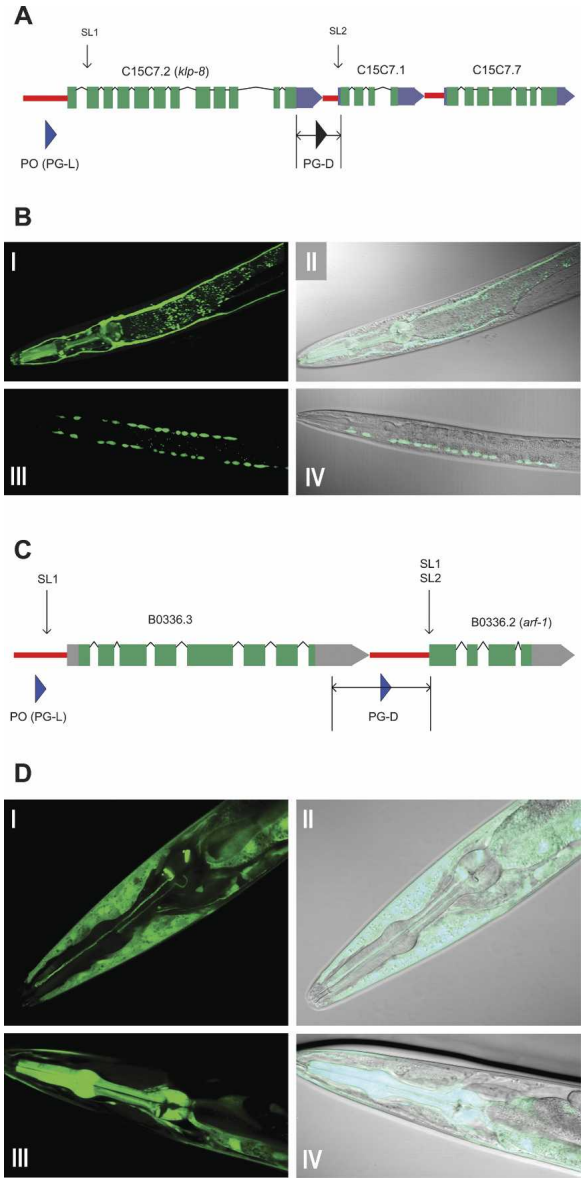

Figure 1. Schematic diagrams and expression patterns of the CEOPX040 (*klp-8*) operon and the CEOP3332 operon. (*A*) The *klp-8* operon. Both the operon promoter (PG-L) for the leading gene C15C7.2 and the internal promoter (PG-D) for the downstream gene C15C7.1 have been extracted. (Green rectangles) Exons; (thin lines) introns; (thick red lines) intergenic regions; (gray areas) UTRs. (PO) The promoter of an operon; (PG-L) a promoter::GFP constructed from the leading gene of an operon; (PG-D) a promoter::GFP constructed from the corresponding operon downstream gene. (*B*) I and II show that the promoter extracted from leading gene C15C7.2 is able to drive GFP expression in the excretory cell, pharynx, pharyngeal neurons, and head neurons, while the promoter extracted from downstream gene C15C7.1 directs GFP expression in the seam cells (III and IV). (*C*) The CEOP3332 operon. Both the operon promoter and the internal promoter for downstream gene B0336.2 are used to construct GFP fusions. (*D*) I and II show that the promoter extracted from leading gene B0336.3 is able to drive GFP expression in hypodermis, pharyngeal gland cell, gut, and nerve ring, while that of downstream gene B0336.2 is able to drive GFP expression in the pharynx, gut, and head neurons (III and IV).

ers, 20.4% use a mixture of SL1/SL2 and 4% use SL1 only for *trans*-splicing. Spliced leaders for 10.5% of them have not been detected (Table1). It is expected that more spliced leaders will be found among downstream genes when more EST sequences or SL data are obtained (Hwang et al. 2004).

In order to determine the relationship between gene duplication and the presence of internal promoters, we compared our data set against the duplicated gene pairs identified by Lynch and Conery (2000). We found that out of 66 downstream genes with internal promoters, 12.1% had paralogous genes, significantly different from those downstream genes without evidence for internal promoters, where only 2.9% belonged to duplicated gene pairs ($P < 0.01$, two-tailed Fisher exact test). In addition, the downstream genes without internal promoters have significantly fewer duplicated gene pairs when compared to the genome as a whole, where 2298 (12.5%) genes belong to duplicated gene pairs ($P < 0.00005$, two-tailed Fisher exact test) (Lynch and Conery 2000).

In order to determine the influence of operon structure on the development of alternative splicing variants, we examined the profiles of alternative splicing of operon downstream genes and of genes in the entire *C. elegans* genome. Surprisingly, although some have alternative 5′ starting sites or 3′ polyadenylation sites, none of the 66 downstream genes with internal promoters have alternatively spliced variants, which is significantly less than the genomic average of 10% ($P < 0.025$, two-tailed χ^2 test). Conversely, 21.5% of the downstream genes without internal promoters are alternatively spliced, a significantly higher proportion than the genomic average ($P < 0.0001$, two-tailed χ^2) and also higher than downstream genes with internal promoters ($P < 0.0001$, two-tailed χ^2). We evaluated whether this difference was due to a lack of cDNA coverage for operon downstream genes with internal promoters, since alternative splicing variants are typically determined through the analysis of EST sequences. The results show that all downstream genes with internal promoters, except T19A5.5, have cDNA coverage (Supplemental Table 1). Figure 2 displays the cDNA coverage for downstream genes with or without internal promoters and for genes in the entire *C. elegans* genome. The average cDNA coverage is ~59 (median 19) for the downstream genes with internal promoters, while the average is ~17 (median 11) for those without internal promoters and 15 (median 5) for genes in the entire genome. Thus, as the

Table 1. Comparison of properties between operon downstream genes with and without internal promoters

	Operon downstream genes with internal promoters	Operon downstream genes without internal promoters
Number of gene-specific promoter::GFPs constructed[a]	66	172
Spliced leader[b]		
None	6%	10.5%
SL1 only	0%	4%
SL2 only	56%	65%
SL1 and SL2	38%	20.4%
With cDNA coverage	98.5%	98.3%
With duplicated gene pairs	12.1%	2.9%
Alternative splice variants	0%	21.5%
Average intergenic spacing (± SE)	596 ± 82 bp	428 ± 36 bp

[a]Studied in this work only.
[b]As annotated in WormBase, based on current evidence, which may be incomplete.

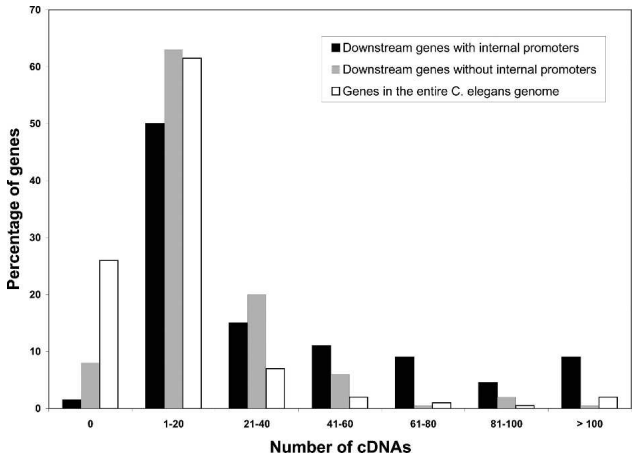

Figure 2. Gene cDNA coverage. Percent of genes with corresponding sequenced ESTs or full-length cDNAs.

cDNA coverage for the operon downstream genes with internal promoters is high, it becomes even more significant that no alternative splicing variants for these genes have been discovered.

Intergenic spacing in putative hybrid operons

The average length of intergenic regions with promoter activity is 596 bp (median 377 bp), while that of the intergenic regions without promoter activity is shorter, around 428 bp with a median of 207 bp ($P < 0.05$, two-tailed *t*-test; Table 1). 53% of the intergenic regions with promoter activities fall within the 0~400-bp range, which agrees with the overall distribution of gene spacing within operons described by Zorio et al. (1994). Overall, however, the hybrid operons studied exhibit a much broader distribution of intergenic spaces with promoter activity (from 56 to 3844 bp; see Supplemental Fig. 2) than that observed by Zorio et al. Therefore, the wide distribution of gene spacing in hybrid operons might reflect the need to accommodate internal promoters and other *cis*-regulatory elements. Our calculations show that 20% of intergenic regions with internal promoter activity have lengths >1000 bp. It is possible that some annotated operons with lengthy intergenic regions might not be true operons. For instance, it is likely that the *mrp-1* gene and the *mrp-2* gene should not be grouped in an operon (CEOPX154), because *mrp-2* is not *trans*-spliced to SL2 and the intergenic distance between *mrp-1* and *mrp-2* is >3800 bp.

Gene expression analysis

SAGE can be used for measuring and comparing gene expression at the RNA level quantitatively. Using SAGE data, we evaluated gene expression profiles for all 65 hybrid operons we have identified, and the 150 operons from WormBase annotation that were not determined to be hybrid (hereafter referred to as non-hybrid operons).

Table 2 shows gene expression correlation across different categories. In general, genes in operons display coexpression when evaluated against comparisons between non-operon genes ($P < 0.001$, two-tailed *t*-test), which agrees with the results of Lercher et al. (2003). The same is true for genes grouped in the highly expressed gene set as defined in Methods. However, the coexpression of genes in operons is weak, although coexpression increases for the highly expressed gene set. Statistically, there is no significant difference in coexpression between genes in hybrid and those in non-hybrid operons. However, this comparison is complicated by the fact that some non-hybrid operons contain more than two genes, and it is not known whether other untested downstream genes in these operons have their own internal promoters.

In order to examine in more detail the impact of internal promoters on gene expression, we analyzed the expression profile for all hybrid operons. Our results showed that gene members in seven hybrid operons were strongly coexpressed ($r \geq 0.5$), while those in the remaining 58 hybrid operons lacked any strong expression correlation. To dissect the reasons for the lack of gene expression correlation in hybrid operons, we also examined differences in absolute expression level in various tissues using the Audic-Claverie algorithm (Audic and Claverie 1997). We found that the lack of strongly correlated gene expression in 25 out of 58 hybrid operons could be explained by the presence of internal promoters. In these cases, the expression level for each downstream gene with an internal promoter showed either no significant difference or was significantly higher when compared to the expression level of its corresponding leading gene in the same hybrid operon. For instance, the expression correlation coefficient for two genes, F40F8.9 (*lsm-1*) and F40F8.1, in the *lsm-1* hybrid operon (CEOP2520) is 0.003 for the tissues and developmental stages tested. Although there is no significant difference in their expression in embryos, larvae, adults, dauer larvae, or most tissue-specific libraries, the amount of F40F8.1 mRNA is statistically higher ($P < 0.05$, Audic-Claverie algorithm; Audic and Claverie 1997) in neural-specific tissues relative to the expression of its upstream gene *lsm-1* (Supplemental Fig. 3). The GFP fluorescence patterns also confirm that the promoter extracted from F40F8.1 is able to drive the reporter expression in neurons (data not shown). This reflects that F40F8.1 expression is not only influenced by the operon promoter, but is additionally controlled by its own internal promoter in certain tissues.

In some cases, issues such as differences in mRNA half-life, lack of SAGE tags for some genes, and ambiguous SAGE tags for others could affect the gene expression analysis and may contribute to a lack of expression correlation. We found that some ambiguous SAGE tags resulted from the prediction of full-length transcripts based on cDNAs covering more than one gene in an

Table 2. Gene coexpression analysis (correlation coefficient of SAGE data *r* ± standard error) for genes in hybrid operons, non-hybrid operons, operons in general, and random gene pairs

	Number of operons[a]	Number of genes[a]	*r*
Hybrid operons			
All	58	170	0.139 ± 0.020
Highly expressed[b]	10	23	0.367 ± 0.076
Non-hybrid operons[c]			
All	118	331	0.143 ± 0.015
Highly expressed	7	18	0.345 ± 0.052
Operons			
All	888	2234	0.144 ± 0.006
Highly expressed	97	210	0.292 ± 0.026
Non-operons			
All (random pairs)	—	12,571	0.019 ± 0.00003
Highly expressed	—	3884	0.206 ± 0.0002

[a]Only includes those with available expression information.
[b]As defined in Methods.
[c]WormBase annotated operons that gave negative GFP results.

operon. One example is the *mai-1* gene and the *gpd-2* gene in the *mai-1* operon. According to WormBase annotations, cDNA yk1412f04.5 covers both *mai-1* and *gpd-2* and causes the *gpd-2* transcript to be annotated as completely overlapping the *mai-1* transcript. According to the works of Spieth et al. (1993) and Liu et al. (2001), however, the 3′-end of *mai-1* is cleaved ~110 bp upstream of *gpd-2*. The same is true for cDNA U24189. Therefore, cDNAs that cover more than one gene in an operon cannot be used for full-length transcript prediction.

Discussion

Our study, the largest analysis of operon promoter activity to date, has shown that 66 out of 238 GFP fusions constructed from operon downstream genes are able to autonomously drive GFP expression in vivo. This indicates the presence of internal promoters in more than one-quarter of the operon downstream genes studied. Both GFP expression patterns and SAGE data demonstrate a complex gene expression regulation system in the hybrid operons of *C. elegans*, where internal promoters are able to drive tissue-specific expression in a pattern distinct from the operon promoter. The functional activity of internal promoters is further demonstrated by the ability of an internal promoter construct to sufficiently recapitulate normal expression to the extent that a lethal phenotype can be rescued. In addition, the GFP promoter assay is able to distinguish two classes of downstream operon genes that differ significantly in their intergenic distance, frequency of paralogs, and the rate at which they undergo alternative splicing, providing further evidence that this assay reflects bona fide promoter activity.

Gene duplication

Gene duplication is one of the most influential processes for the evolution of genetic diversity (Ohno 1970). According to Rubin et al. (2000), gene duplications affect almost half of the annotated genes in the *C. elegans* genome. The number is reduced to 2298 (12.5%) if using more stringent measurements (Lynch and Conery 2000). Our data indicate that genes lacking their own promoter in an operon are less likely to have been involved in a recent duplication event ($P < 0.01$), presumably because such genes are unlikely to be successfully transcribed and retained unless the entire operon undergoes duplication (Lercher et al. 2003). Therefore, it is possible that internal promoters facilitate the duplication of their corresponding downstream genes within operons to other genomic locations. Alternatively, it is also possible that genes outside operons duplicate into an existing operon or beside another gene to form a new operon.

Alternative splicing variants

Besides gene duplication, the process of alternative splicing also provides an important way for increasing the functional diversity of a gene (Kriventseva et al. 2003). According to WormBase annotations, ~10% of the annotated genes in the *C. elegans* genome are alternatively spliced. Hence, it is intriguing that none of the operon downstream genes with internal promoters have alternative splicing variants, which is significant when compared to the genome as a whole ($P < 0.025$) and even more significant when compared to the increased number of alternative splicing variants seen for downstream genes without internal promoters ($P < 0.0001$). One explanation for the increased alternative splicing in operon genes lacking their own promoters might be that since they are more limited in their ability to successfully duplicate, novel functionality can only be evolved through the creation of novel isoforms. Conversely, not only can downstream genes with internal promoters duplicate readily, but they can be regulated through either the operon promoters or their own internal promoters. Therefore, we find that genes in operons with internal promoters represent the first class of multi-exonic genes in *C. elegans* that appear to be constrained in their ability to produce alternatively spliced variants. This class might be anticipated to provide useful insight in studies of the mechanisms underlying alternative splicing.

SL1 and SL2

Since most operon pre-mRNAs are *trans*-spliced cotranscriptionally, it has proven to be very difficult to experimentally detect polycistronic pre-mRNAs. The presence of SL2 *trans*-splicing sites among closely spaced genes with the same orientation has become the major determinant for the prediction of operon downstream genes. Zhao et al. (2004), however, questioned the role of SL2 and suggested that the presence of SL2 might not be sufficient for downstream gene prediction in *C. elegans* operons.

Do patterns of *trans*-splicing in downstream genes signal the presence or absence of internal promoters? Our analysis shows that nearly half of the operon downstream genes with internal promoters studied here only have evidence for *trans*-splicing to SL2. We might expect that proportion to be reduced when more SL1 and SL2 data are obtained (Hwang et al. 2004). However, since 30% of genes receive no *trans*-splice sequence at all (Zorio et al. 1994), it is unlikely that we will ever be assured of capturing all possible spliced leaders. Therefore, the absence of a mixture of SL1 and SL2 does not necessarily indicate that there are no internal promoters for the downstream genes. Both Spieth et al. (1993) and Liu et al. (2001) observed that the failure of the 3′-end cleavage of upstream genes would cause the *trans*-splice machinery to treat the downstream genes in the same fashion as the upstream ones. In this case, the downstream genes are *trans*-spliced to SL1 as well as SL2. Furthermore, Spieth et al. (1993) indicate that the *gpd-3* gene in the *mai-1* operon does not have its own internal promoter, while the *gpd-3* gene is *trans*-spliced to both SL1 and SL2 according to WormBase. Therefore, it follows that *trans*-splicing with a mixture of SL1 and SL2 is not a sufficient indicator for the prediction of an internal promoter.

Interestingly, in the *C. elegans* genomic annotation (WS135) analyzed, >50 genes are indicated to be *trans*-spliced to either a mixture of SL1 and SL2 or exclusively SL2 despite having no evidence that they are downstream participants in operons (data not shown). This may suggest that there exists an alternative mechanism for SL2 *trans*-splicing. Conversely, this may indicate the presence of undiscovered genes in these regions that would function as the initiating gene in an undefined operon. Since the *C. elegans* genomic annotation is relatively mature, it is appealing to speculate that these undiscovered genes could be non-protein coding.

Approximately 20% of all *C. elegans* operons were subjected to GFP analysis in this study. Although our data set does not cover every internal promoter in the annotated operons of the *C. elegans* genome, it greatly enhances our knowledge and understanding of *C. elegans* promoters, especially as operon downstream genes were specifically excluded in previous large-scale promoter studies in *C. elegans* (McKay et al. 2003; Dupuy et al. 2004). Our work, consistent with previous findings in the

literature, suggests that the presence of a mixture of SL1 and SL2 is not indicative of internal promoters. Additionally, we find significant differences in the occurrence of alternative splicing variants and gene duplications related to the presence or absence of internal promoters. This may reflect differences in selective pressures, as genes with internal promoters are better able to undergo gene duplication, while genes in operons without internal promoters are more limited because of lack of promoters and *cis*-regulatory elements, and thus may be forced to develop alternative splicing variants to adapt during evolution. Hybrid operons in *C. elegans* represent an alternative genomic structure that provides further flexibility in gene organization and expression and may be an influential factor in gene evolution.

Methods

Discovery of operons with putative internal promoters in *C. elegans*

All described gene and operon annotations and genomic sequences of *C. elegans* were extracted from WormBase release WS135 (ftp.sanger.ac.uk).

The directionally stitched promoter::GFP fusions were constructed according to McKay et al. (2003), available from http://elegans.bcgsc.bc.ca/promoter_primers. The results of promoter::GFP experiments are available from http://elegans.bcgsc.bc.ca/perl/eprofile. The observed expression is unlikely to occur by chance since, using a promoter trapping technique, Hope (1991) observed that <1% of the random DNA segments (5~10 kb) from the *C. elegans* genome acted as promoters able to drive lacZ expression. Among those positive ones, he and his colleagues were able to find some true promoters for previously uncharacterized genes (Hope 1994; Hope et al. 1998).

Gene-specific promoter::GFP constructs with positive reporter expression were identified. Only those genes residing downstream in WormBase annotated operons were used for further analysis. We then mapped the coordinates of primer sets used in GFP fusions through the e-PCR program (Schuler 1997) to verify that each primer set targeted the intended region, excluding those that overlapped with the corresponding operon promoter at the 5′-end of the leading gene. The word size for the e-PCR program was set at 7 and the mismatch at 2. Constructs and their corresponding genes with no observed reporter expression were also obtained for comparison.

Intergenic distance within an operon was defined as the region between the end of the poly(A) tail of the upstream gene and the beginning of the *trans*-spliced site of the corresponding downstream gene.

Paralogous genes and annotated spliced leaders in the *C. elegans* genome

Paralogous gene pairs (or duplicated gene pairs) in *C. elegans* were obtained from Lynch and Conery (2000). The presence of SL2 was determined either through the examination of EST sequences or from the results of SL2-primed microarray experiments (Blumenthal et al. 2002) as annotated by WormBase. Also included in the SL2 category are the SL2-like SL3 and SL4 sequences, which are regarded to be functionally equivalent to SL2 (Ross et al. 1995; Hough et al. 1999; see Supplemental Table 1).

C. elegans genes with alternative splicing variants

A total of 1870 genes with 4489 alternative splicing variants were extracted from WormBase release WS135 (ftp.sanger.ac.uk). Genes belonging to operon downstream units and used for GFP fusions were identified. We manually examined the 5′- and 3′-ends of these genes to ensure they were aligned correctly with their corresponding cDNAs. In several cases we noted internal SL1 attachment sites, which would result in shortened final transcripts missing sequence at the 5′-end. However, these were not annotated as different transcripts by WormBase and were not considered as alternative splice variants. These are likely to represent rare, incorrectly *trans*-spliced transcript forms that may occur because of the highly similar nature of the consensus sequences for both *trans*- and *cis*-splicing sites in *C. elegans* (Blumenthal 1995). For example, the transcript generated from the internal SL1 site for the Y57G11C.11 (*coq-3*) gene would produce a truncated protein product.

Gene expression analysis

A total of 26 SAGE (serial analysis of gene expression; Velculescu et al. 1995) libraries were used and normalized to 100,000 tags, after excluding those libraries with original library sizes <60,000 tags (http://tock.bcgsc.bc.ca/cgi-bin/sage; McKay et al. 2003). Among those selected, nine were LongSAGE (Saha et al. 2002) libraries (21 bp), while the remaining 17 were short SAGE libraries (14 bp).

SAGE tags were mapped to genes using the full-length transcripts predicted from WormBase release WS110 based on a previously described method (Pleasance et al. 2003). All duplicate ditags and unmatched and antisense tags were removed. Therefore, only unambiguously mapped SAGE tags were used. The minimal tag count was set to 1 with the SAGE tag sequencing error rate <1%. Different SAGE tags were combined if they targeted the same gene or they belonged to alternative splicing variants of the same gene.

The Pearson correlation coefficient was used for gene expression correlation analysis, while the Audic-Claverie algorithm (Audic and Claverie 1997) was used to determine the statistical significance of a differential change of gene expression at the 95% confidence level. For a pairwise coexpression calculation, genes were considered highly expressed if every gene in a comparison had a normalized SAGE tag count ≥10 per 100,000 in at least one of the same libraries. Genes within an operon were considered coexpressed strongly if the correlation coefficient was ≥0.5.

Transgenic complementation of *let–721* (C05D11.12)

The lethal phenotype of both alleles of *let-721*, *s2447* and *s2812*, was complemented by transgenic rescue with the cosmid C05D11 (GenBank U00048) and the plasmid pC05.12, which is a subclone of C05D11 that contains only C05D11.12 and the flanking 5′ and 3′ regions (bases 41,634–45,272 of C05D11). Both C05D11 and pC05.12 have 1137 bp of sequence upstream of the translation start site of C05D11.12 and do not contain any of the upstream genes in the operon CEOP3384. C05D11 and pC05.12 were each injected at 5 ng/μL together with 95 ng/μL *pCes1943[rol-6(su1006)dm]*. Stable transgenics were mated to the strains BC4157 *sDp3(III;f);dpy-17(e164) let-721(s2447) unc-32(e189)* or BC4842 *sDp3(III;f);dpy-17(e164) let-721(s2812) ncl-1(e1865) unc-32(e189)*, and F2 Dpy Unc animals were scored for viability.

Deletion analysis

Many factors might contribute to the failure of GFP expression. For example, the promoter may be expressed only in male worms or at the dauer stage, the reporter GFP may be expressed at a very low level, or we may fail to extract the entire promoter sequence.

To address some of these concerns, we conducted the following experiment. From the promoter::GFP results, we selected a set of 12 operons, where the promoters of upstream genes presented positive gene expression patterns while the first corresponding downstream genes gave negative results. In order to evaluate gene expression modulatory effects produced via the intra-operonic region between the two genes, we designed the left primer right after the start codon of the upstream gene with the upstream operon promoter deleted, while still using the same right primer used for the negative downstream gene. Two control upstream genes (K12H4.5 in the CEOP3476 operon and Y71F9B.3 in the CEOP1056 operon) from the above operon set were tested again, where the left primers were the same as those used in the positive upstream genes and the right primers were the same as those used for the corresponding negative downstream genes.

Acknowledgments

We thank Thomas Blumenthal for critical review and comments on the manuscript. We thank Richard Durbin, Anthony Rogers, and Daniel Lawson of WormBase at the Sanger Institute for information and consultation regarding operon annotations and full-length transcript prediction. We also thank Courtney Mills for providing the analytical Web sites for both SAGE and GFP experiments. E.P. is supported by the Canadian Institutes of Health Research and the Michael Smith Foundation for Health Research (MSFHR). A.M. is supported by a NSERC graduate scholarship. D.L.B. holds a Canada Research Chair. The work was supported in part by a CIHR grant to D.L.B. S.J. and M.M. are Scholars of the MSFHR. This work was primarily funded by Genome Canada.

References

Andrews, J., Smith, M., Merakovsky, J., Coulson, M., Hannan, F., and Kelly, L.E. 1996. The stoned locus of *Drosophila melanogaster* produces a dicistronic transcript and encodes two distinct polypeptides. *Genetics* **143:** 1699–1711.

Audic, S. and Claverie, J.M. 1997. The significance of digital gene expression profiles. *Genome Res.* **7:** 986–995.

Blumenthal, T. 1995. *Trans*-splicing and polycistronic transcription in *Caenorhabditis elegans*. *Trends Genet.* **11:** 132–136.

Blumenthal, T. 2004. Operons in eukaryotes. *Brief. Funct. Genomic. Proteomic.* **3:** 199–211.

Blumenthal, T. 2005. *Trans*-splicing and operons. In *WormBook* (ed. The *C. elegans* Research Community). WormBook. http://www.wormbook.org.

Blumenthal, T. and Gleason, K.S. 2003. *Caenorhabditis elegans* operons: Form and function. *Nat. Rev. Genet.* **4:** 112–120.

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.

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W., and Prasher, D.C. 1994. Green fluorescent protein as a marker for gene expression. *Science* **263:** 802–805.

Chang, S., Johnston Jr., R.J., and Hobert, O. 2003. A transcriptional regulatory cascade that controls left/right asymmetry in chemosensory neurons of *C. elegans*. *Genes & Dev.* **17:** 2123–2137.

Corcoran, M.M., Hammarsund, M., Zhu, C., Lerner, M., Kapanadze, B., Wilson, B., Larsson, C., Forsberg, L., Ibbotson, R.E., Einhorn, S., et al. 2004. DLEU2 encodes an antisense RNA for the putative bicistronic RFP2/LEU5 gene in humans and mouse. *Genes Chromosomes Cancer* **40:** 285–297.

Dupuy, D., Li, Q.R., Deplancke, B., Boxem, M., Hao, T., Lamesch, P., Sequerra, R., Bosak, S., Doucette-Stamm, L., Hope, I.A., et al. 2004. A first version of the *Caenorhabditis elegans* Promoterome. *Genome Res.* **14:** 2169–2175.

Evans, D. and Blumenthal, T. 2000. *trans* splicing of polycistronic *Caenorhabditis elegans* pre-mRNAs: Analysis of the SL2 RNA. *Mol. Cell. Biol.* **20:** 6659–6667.

Garcia-Rios, M., Fujita, T., LaRosa, P.C., Locy, R.D., Clithero, J.M., Bressan, R.A., and Csonka, L.N. 1997. Cloning of a polycistronic cDNA from tomato encoding γ-glutamyl kinase and gamma-glutamyl phosphate reductase. *Proc. Natl. Acad. Sci.* **94:** 8249–8254.

Hope, I.A. 1991. 'Promoter trapping' in *Caenorhabditis elegans*. *Development* **113:** 399–408.

Hope, I.A. 1994. PES-1 is expressed during early embryogenesis in *Caenorhabditis elegans* and has homology to the fork head family of transcription factors. *Development* **120:** 505–514.

Hope, I.A., Arnold, J.M., McCarroll, D., Jun, G., Krupa, A.P., and Herbert, R. 1998. Promoter trapping identifies real genes in *C. elegans*. *Mol. Gen. Genet.* **260:** 300–308.

Horowitz, H. and Platt, T. 1983. Initiation in vivo at the internal trp p2 promoter of *Escherichia coli*. *J. Biol. Chem.* **258:** 7890–7893.

Hough, R.F., Lingam, A.T., and Bass, B.L. 1999. *Caenorhabditis elegans* mRNAs that encode a protein similar to ADARs derive from an operon containing six genes. *Nucleic Acids Res.* **27:** 3424–3432.

Huang, X.Y. and Hirsh, D. 1989. A second *trans*-spliced RNA leader sequence in the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci.* **86:** 8640–8644.

Hwang, B.J., Muller, H.M., and Sternberg, P.W. 2004. Genome annotation by high-throughput 5′ RNA end determination. *Proc. Natl. Acad. Sci.* **101:** 1650–1655.

Imboden, M.A., Laird, P.W., Affolter, M., and Seebeck, T. 1987. Transcription of the intergenic regions of the tubulin gene cluster of *Trypanosoma brucei*: Evidence for a polycistronic transcription unit in a eukaryote. *Nucleic Acids Res.* **15:** 7357–7368.

Jacob, F. and Monod, J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3:** 318–356.

Kriventseva, E.V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M.S., and Sunyaev, S. 2003. Increase of functional diversity by alternative splicing. *Trends Genet.* **19:** 124–128.

Langer, D., Hain, J., Thuriaux, P., and Zillig, W. 1995. Transcription in archaea: Similarity to that in eucarya. *Proc. Natl. Acad. Sci.* **92:** 5768–5772.

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.

Liu, H., Jang, J.K., Graham, J., Nycz, K., and McKim, K.S. 2000. Two genes required for meiotic recombination in *Drosophila* are expressed from a dicistronic message. *Genetics* **154:** 1735–1746.

Liu, Y., Huang, T., MacMorris, M., and Blumenthal, T. 2001. Interplay between AAUAAA and the *trans*-splice site in processing of a *Caenorhabditis elegans* operon pre-mRNA. *RNA* **7:** 176–181.

Lu, C., Bentley, W.E., and Rao, G. 2004. A high-throughput approach to promoter study using green fluorescent protein. *Biotechnol. Prog.* **20:** 1634–1640.

Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.

Macejak, D.G. and Sarnow, P. 1991. Internal initiation of translation mediated by the 5′ leader of a cellular mRNA. *Nature* **353:** 90–94.

MacMorris, M., Kumar, M., Lasda, E., Larsen, A., Kraemer, B., and Blumenthal, T. 2007. A novel family of *C. elegans* snRNPs contains proteins associated with *trans*-splicing. *RNA* **13:** 511–520.

McKay, S.J., Johnsen, R., Khattra, J., Asano, J., Baillie, D.L., Chan, S., Dube, N., Fang, L., Goszczynski, B., Ha, E., et al. 2003. Gene expression profiling of cells, tissues, and developmental stages of the nematode *C. elegans*. *Cold Spring Harb. Symp. Quant. Biol.* **68:** 159–169.

Murphy, W.J., Watkins, K.P., and Agabian, N. 1986. Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: Evidence for *trans* splicing. *Cell* **47:** 517–525.

Nilsen, T.W. 2001. Evolutionary origin of SL-addition *trans*-splicing: Still an enigma. *Trends Genet.* **17:** 678–680.

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, New York.

Pflugrad, A., Meir, J.Y., Barnes, T.M., and Miller III, D.M. 1997. The Groucho-like transcription factor UNC-37 functions with the neural specificity gene *unc-4* to govern motor neuron identity in *C. elegans*. *Development* **124:** 1699–1709.

Pleasance, E.D., Marra, M.A., and Jones, S.J. 2003. Assessment of SAGE in transcript identification. *Genome Res.* **13:** 1203–1215.

Ross, L.H., Freedman, J.H., and Rubin, C.S. 1995. Structure and expression of novel spliced leader RNA genes in *Caenorhabditis elegans*. *J. Biol. Chem.* **270:** 22066–22075.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., et al. 2000. Comparative genomics of the eukaryotes. *Science* **287:** 2204–2215.

Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B.,

Kinzler, K.W., and Velculescu, V.E. 2002. Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20:** 508–512.
Saldi, T., Wilusz, C., Macmorris, M., and Blumenthal, T. 2007. Functional redundancy of worm spliceosomal proteins U1A and U2B. *Proc. Natl. Acad. Sci.* **104:** 9753–9757.
Schuler, G.D. 1997. Sequence mapping by electronic PCR. *Genome Res.* **7:** 541–550.
Semple, C. and Wolfe, K.H. 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. *J. Mol. Evol.* **48:** 555–564.
Singer, G.A., Lloyd, A.T., Huminiecki, L.B., and Wolfe, K.H. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol. Biol. Evol.* **22:** 767–775.
Sloan, J., Kinghorn, J.R., and Unkles, S.E. 1999. The two subunits of human molybdopterin synthase: Evidence for a bicistronic messenger RNA with overlapping reading frames. *Nucleic Acids Res.* **27:** 854–858.
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.
Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J. 2001. WormBase: Network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.* **29:** 82–86.
Thimmapuram, J., Duan, H., Liu, L., and Schuler, M.A. 2005. Bicistronic and fused monocistronic transcripts are derived from adjacent loci in the *Arabidopsis* genome. *RNA* **11:** 128–138.
Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P., and Myers, R.M. 2004. An abundance of bidirectional promoters in the human genome. *Genome Res.* **14:** 62–66.
Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.
Ye, X., Fong, P., Iizuka, N., Choate, D., and Cavener, D.R. 1997. Ultrabithorax and Antennapedia 5′ untranslated regions promote developmentally regulated internal translation initiation. *Mol. Cell. Biol.* **17:** 1714–1721.
Zhao, Z., Sheps, J.A., Ling, V., Fang, L.L., and Baillie, D.L. 2004. Expression analysis of ABC transporters reveals differential functions of tandemly duplicated genes in *Caenorhabditis elegans*. *J. Mol. Biol.* **344:** 409–417.
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.

Received June 18, 2007; accepted in revised form June 29, 2007.