A global analysis of *C. elegans trans*-splicing

Mary Ann Allen, LaDeana W. Hillier, Robert H. Waterston, et al.

Genome Res. 2011 21: 255-264 originally published online December 22, 2010
Access the most recent version at doi:10.1101/gr.113811.110

A global analysis of *C. elegans trans*-splicing

Mary Ann Allen,[1] LaDeana W. Hillier,[2] Robert H. Waterston,[2] and Thomas Blumenthal[1,3]

[1]*Department of Molecular, Cellular, and Developmental Biology, University of Colorado at Boulder, Colorado 80309, USA;*
[2]*Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195-5065, USA*

Trans-splicing of one of two short leader RNAs, SL1 or SL2, occurs at the 5′ ends of pre-mRNAs of many *C. elegans* genes. We have exploited RNA-sequencing data from the modENCODE project to analyze the transcriptome of *C. elegans* for patterns of *trans*-splicing. Transcripts of ~70% of genes are *trans*-spliced, similar to earlier estimates based on analysis of far fewer genes. The mRNAs of most *trans*-spliced genes are spliced to either SL1 or SL2, but most genes are not *trans*-spliced to both, indicating that SL1 and SL2 *trans*-splicing use different underlying mechanisms. SL2 *trans*-splicing occurs in order to separate the products of genes in operons genome wide. Shorter intercistronic distance is associated with greater use of SL2. Finally, increased use of SL1 *trans*-splicing to downstream operon genes can indicate the presence of an extra promoter in the intercistronic region, creating what has been termed a "hybrid" operon. Within hybrid operons the presence of the two promoters results in the use of the two SL classes: Transcription that originates at the promoter upstream of another gene creates a polycistronic pre-mRNA that receives SL2, whereas transcription that originates at the internal promoter creates transcripts that receive SL1. Overall, our data demonstrate that >17% of all *C. elegans* genes are in operons.

[Supplemental material is available for this article. The sequence data from this study have been submitted to the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession nos. SRA008646 and SRA003622.]

C. elegans uses two RNA processing features that distinguish it from other model organisms. First, the transcripts of many genes are *trans*-spliced to a spliced leader (SL). *Trans*-splicing is a process in which an SL replaces the 5′ end of a transcript by spliceosomal splicing. The 22-nucleotide (nt) SL is donated by an ~100-nt SL snRNP (small nuclear ribonucleoprotein) to a pre-mRNA with an intron-like region, the outron, containing an unpaired 3′ splice site located near the 5′ end. The second distinguishing feature is that many genes are transcribed in polycistronic units, known as operons, where a single promoter serves several genes. The operons can be up to eight genes long, and the polycistronic pre-mRNAs are separated into individual cistrons by 3′ end formation accompanied by SL *trans*-splicing.

These two features have important implications for *C. elegans* research. For instance, deletions/insertions within an operon may affect not only the expression of the gene containing the mutation, but also the genes downstream from it in the operon (Cui et al. 2008). Similarly, in a strain with enhanced RNAi sensitivity, RNAi of an operon gene can also affect expression of genes downstream (Guang et al. 2010). Finally, the *trans*-splice site is at the 5′ end of the mRNA, not the pre-mRNA, and thus, the promoter is often not directly adjacent to the gene, but rather upstream of the outron or the entire operon (Blumenthal and Spieth 1996).

SL *trans*-splicing has been reported in many phyla, including trypanosomes, nematodes, and even chordates (Sutton and Boothroyd 1986; Krause and Hirsh 1987; Vandenberghe et al. 2001). In 1994, it was estimated that 70% of *C. elegans* genes were *trans*-spliced, based on the limited genomic and cDNA sequence data available (Zorio et al. 1994). This estimate was based on both analysis of 37 cosmids for 3′ splices sites that might be able to act as *trans*-splice sites, and a survey of genes from the literature. Now, the use of new sequencing technologies should allow us to estimate the percentage of genes *trans*-spliced much more accurately.

[3]Corresponding author.
E-mail tom.blumenthal@colorado.edu.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.113811.110.

The donor in *trans*-splicing is a small nuclear ribonucleoprotein called the SL snRNP. *C. elegans* uses two types of SL snRNP: SL1 and SL2, the latter of which has several sequence variants, termed SL3–SL12 (Blumenthal 2005; MacMorris et al. 2007). SL1 is thought to *trans*-splice both to non-operon genes and to first genes in operons, thereby removing the outron (Conrad et al. 1991). SL2 is thought to *trans*-splice to downstream genes in operons. In fact, the *C. elegans* operons were discovered as a result of this specialized *trans*-splicing to genes in the same orientation in tightly packed clusters (Spieth et al. 1993). When such genes were analyzed, they all were found to be *trans*-spliced to SL2. Conversely, when genes that had been shown to be SL2 *trans*-spliced were analyzed, they were subsequently found to be downstream in tightly linked gene clusters. Several polycistronic cDNAs from RNAs that had not yet been processed were identified (Zorio et al. 1994). Subsequently, the overwhelming correlation between these clusters and SL2 *trans*-splicing to downstream genes was demonstrated by microarray analysis (Blumenthal and Spieth 1996).

The current method of operon annotation involves manual curation based on both SL2 *trans*-splicing and a short distance between genes on the same strand. However, several genes fall into categories that make it difficult to determine whether they are within an operon. For instance, some genes are SL2 *trans*-spliced but have a long distance to the next gene upstream, whereas others are mostly SL1 *trans*-spliced but have a short distance to the next gene upstream. Furthermore, a few genes receive a mixture of SL1 and SL2. In addition, as many as 25% of operons were estimated to be "hybrid operons" in which there are internal promoters (Huang et al. 2007). Hybrid operons, therefore, add an additional layer of complexity to the annotation of operons.

The most accurate method for annotation of operons would be a demonstration of polycistronic transcripts, but rapid processing of polycistronic transcripts prevents identification of most operons this way. However, if SL2 is specific for downstream genes in operons, global analysis of SL2 should yield an accurate list of all operons. Furthermore, if SL1 is specific for genes with adjacent promoters, hybrid operons could be identified by *trans*-splice sites that receive a mixture of SL1 and SL2.

As part of the modENCODE project (Celniker et al. 2009), deep RNA-sequence data has been generated for 19 different stages and conditions (Hillier et al. 2009; Gerstein et al. 2010). High-throughput sequence reads in this project (>1 billion total) were mapped to the genome and to candidate splice junctions. Using the 28,249 *trans*-splice sites identified, the estimate of how many genes are *trans*-spliced can be improved. Here it is demonstrated that, on a whole-genome level, genes that receive SL2 are correctly annotated as downstream in operons, since genes that receive high levels of SL2 are located in densely packed gene clusters and, in general, lack proximal promoters. We also demonstrate globally that genes that receive a mixture of SL1/SL2 can be in hybrid operons and that SL2 is used when the transcript is from the operon promoter, whereas SL1 is used when the transcript is from the proximal promoter. In addition, *trans*-splicing events are also documented at 3′ splice sites of long introns.

Results

Approximately 70% of all *C. elegans* genes are *trans*-spliced

To discover which genes are *trans*-spliced, the *trans*-splice sites identified by RNA sequencing to the 5′ ends of annotated genes were mapped (see Methods). Almost all *C. elegans trans*-splice sites are *trans*-spliced more than 90% of the time (Supplemental Fig. 1). A relatively stringent criterion to identify *trans*-spliced genes was used: either the *trans*-splice site or the first start codon (AUG) downstream of the *trans*-splice site needed to be <500 nt from an annotated 5′ end. By these criteria, 14,157 of the 28,249 *trans*-splice sites (Supplemental File 1) mapped to the 5′ end of 11,387 annotated genes (Supplemental File 2; Supplemental Table 1 categorizes *trans*-splice sites). Interestingly, 2130 genes (19%) had more than one *trans*-splice site, and many of these sites are alternative first exons. These results show that the mRNAs from at least 56% of *C. elegans* genes are *trans*-spliced (Fig. 1A).

Originally, 70% of genes were estimated to be *trans*-spliced (Zorio et al. 1994). We detect only 56% percent of genes as *trans*-spliced. However, perhaps not every *trans*-splicing event was detected because of low levels of expression of some genes. To test for this possibility, all genes with expression levels below certain thresholds were removed from the analysis. Depth of coverage per base per million reads (dcpm) for each gene was used as a read-out of expression level (Hillier et al. 2009). When a low dcpm of 0.05 (~1× average coverage/nt) was used, 71% of these genes were found to be *trans*-spliced. Furthermore, the estimate increases to 83% and 84% when the required coverage level increases further (Fig. 1A). Do highly expressed genes have a greater propensity to be *trans*-spliced, or is *trans*-splicing of low expressed genes sometimes below detection levels even at this threshold level?

To distinguish between these possibilities, the *trans*-spliced genes were binned based on their expression levels. The data clearly indicate that the greater the expression of the gene the more likely it is to be *trans*-spliced (Fig. 1B). At an average expression level of >0.1 dcpm (~2× coverage per stage), few *trans*-splicing events would be missed if present, and yet the percent of genes *trans*-spliced continues to increase with increased expression level. More highly expressed genes clearly have a greater propensity to be *trans*-spliced. Therefore the highest cutoff levels used in Figure 1A cannot be used to estimate the number of *trans*-spliced genes. Because a 1× coverage of a gene would likely show all *trans*-splicing events, the best estimate is that ~70% of genes are *trans*-spliced and ~84% of highly expressed genes are *trans*-spliced. It should be noted that our data do not include non-*trans*-spliced reads, so any genes that produce non-*trans*-spliced mRNAs of *trans*-spliced genes would not be revealed.

A.

Minimum gene expression level cutoff	# of Genes	% of Genes Trans-spliced
No cut-off	19876	56%
Low (dcpm>0.05)	14043	71%
Medium (dcpm>0.5)	7344	83%
High (dcpm>5)	858	84%

B.

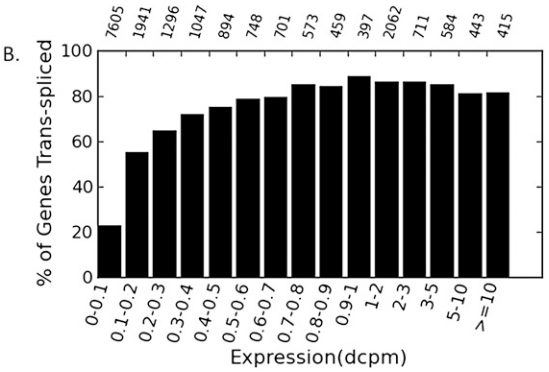

Figure 1. The level of *trans*-splicing in *C. elegans*. (*A*) *Trans*-splice sites were mapped to the 5′ ends of protein-coding genes. For each gene, average expression across all stages in dcpm was used. To guarantee *trans*-splicing could have been detected if present, cut-offs were used to remove genes with low expression levels. The first row contains all genes. The *lower* rows contain genes with expression levels higher than the minimal expression level cutoff. The *middle* column lists the number of genes above the cutoff, and the *last* column is the percentage of those genes that are *trans*-spliced. (*B*) Genes expressed more highly are more likely to be *trans*-spliced. Genes were divided into bins based on expression level. For each gene, average expression across all stages in dcpm was used. Percentage of genes in each group that are *trans*-spliced was plotted. The number *above* each bar indicates the number of genes in the bin.

To determine whether certain kinds of genes are more likely than others of being subject to *trans*-splicing, we compared the functions (Gene Ontology [GO] terms) of *trans*-spliced genes with non-*trans*-spliced genes (Supplemental File 4). Because *trans*-spliced genes, as a group, are expressed at higher levels than non-*trans*-spliced genes, we needed to ensure that any differences we found were not simply a consequence of expression levels. Thus, we compared datasets of genes with GO terms and with a dcpm >0.05, which included 6507 *trans*-spliced and 1842 non-*trans*-spliced genes. These data show that genes involved in development and biological regulation are over-represented in the *trans*-spliced group. In contrast, genes for carbohydrate-binding proteins, phosphate transport, anion transport, and cuticle constituents are over-represented in the non-*trans*-spliced set.

Because it was necessary to calculate the position of the first AUG after each *trans*-splice site for the purpose of categorizing the genes, we used this data to survey the distance between the two sites (Supplemental Fig. 2). In most cases this distance is <10 bp, as was seen with a much smaller data set (Blumenthal and Steward 1997).

SL1 use is common, and SL2 is reserved for genes downstream in operons

Previous results had indicated that SL1 is spliced primarily to outron splice sites near promoters, whereas SL2 and its variants

were primarily spliced to *trans*-splice sites of downstream operon genes. The deep-sequencing data analyzed here show that SL2 and SL2 variants are spliced to the same splice sites, but SL1 is distinct (Supplemental Fig. 3). In order to examine the question of whether sites received only SL1, only SL2, or a mixture of both, the ratio of SL1 to SL2 *trans*-spliced at each position was calculated, and all of the sites were divided into 100 bins from 0% to 100% SL2 (Supplemental Fig. 4A). Because *trans*-splice sites with very few reads do not give an accurate SL1/SL2 ratio, we used the binomial exact test on the percent SL2 and removed any sites that had a *P*-value of more than 0.05. Only the 13,718 *trans*-splice sites with 10 or more reads were analyzed in Supplemental Figure 4B. The data clearly show a peak of >9600 *trans*-splice sites with 100% SL1, and a second peak with mostly SL2. These data strongly support the previous observation that SL1 and SL2 are mechanistically separate and distinct phenomena, since the majority of *trans*-splice sites are *trans*-spliced either to high levels of SL1 or SL2 and far fewer *trans*-splice sites receive a mixture of the two SLs. Interestingly, whereas the SL1 group of genes generally receive 99%–100% SL1, the SL2 genes generally receive only 80%–95% SL2.

The SL1/SL2 ratios of *trans*-splice sites that map to the 5′ ends of genes are given in Figure 2A. The majority of *trans*-spliced genes (82%) are *trans*-spliced to SL1 predominantly (left-most bars), while a smaller group of genes are *trans*-spliced to mostly SL2 (12%). Third, a much smaller group of genes (~6%) are *trans*-spliced to a mixture of SL1 and SL2. To determine whether the genes annotated in WormBase as downstream in operons were primarily SL2 *trans*-spliced, the *trans*-spliced genes were subdivided into three categories: non-operon genes (Fig. 2B), first genes in operons (Fig. 2C), and downstream genes in operons (Fig. 2D). The mRNAs from non-operon genes are *trans*-spliced to SL1 almost exclusively, as are genes located at the 5′ ends of operons that are *trans*-spliced. Interestingly, 85% of first genes in operons are *trans*-spliced. In sharp contrast to the first genes in operons, genes annotated as downstream in operons are *trans*-spliced to high levels of SL2 or sometimes a mixture of SL1 and SL2. Based on the analysis presented here, we have computationally predicted a set of *C. elegans* operons (Supplemental File 3) based solely on downstream genes receiving >10% SL2. The result is that nearly 90% of previously annotated operons overlap with the operons predicted in this way (Supplemental Fig. 5). The operon list created from the deep sequencing data alone contains 3483 genes, >17% of all *C. elegans* genes. Since we no doubt missed some operons due to low expression levels, we estimate that as many as 20% of all *C. elegans* genes are transcribed in operons.

There are ~200 genes previously annotated as downstream in operons whose mRNAs are *trans*-spliced almost entirely to SL1 (Fig. 2D). These genes are likely of two types. The first type encompasses genes in SL1-type operons, where the *trans*-splice site occurs on the same nucleotide as the polyadenylation site of the upstream gene (Williams et al. 1999). We have documented ~30 operons of this type (data not shown). Most of the remainder are genes that appear to have been misannotated as being within operons. We are examining these genes individually to correct the annotation.

The percent of SL2 varies as a function of the distance to the upstream gene

If the distance between the 3′ end of the upstream gene and the *trans*-splice site of the downstream gene is mechanistically important in specifying SL2 as the spliced leader, we might expect there to be an inverse relationship between the intercistronic region (ICR) length and the percent SL2 usage. We calculated the distance to the next gene upstream and plotted these data vs. the percent SL2 on a box and whiskers plot (Fig. 3). The box on the left shows that genes *trans*-spliced mostly to SL1 are in general quite far from the next upstream gene. However, even a very small amount of SL2 *trans*-splicing dramatically lowers the distribution of distances to the next gene upstream. Importantly, the smaller the distance between genes, the higher the percentage of SL2 *trans*-splicing, clearly suggesting a mechanistic relationship between ICR length and SL2 *trans*-splicing. Finally, the single bar to the right of the vertical line shows that when all of the downstream genes in operons are considered together, they are closely apposed to genes upstream.

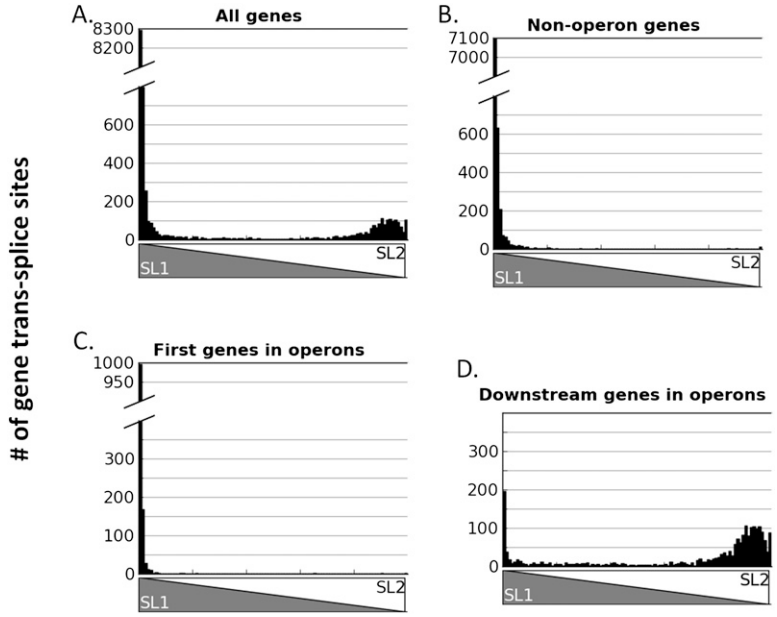

Figure 2. SL1/SL2 *trans*-splicing ratios at each *trans*-spliced gene. Histogram showing the number of gene *trans*-splice sites with indicated SL1:SL2 proportion on the *y*-axis. The *x*-axis shows the proportion SL1/SL2 by the opposing triangles and ranges from 100% SL1 to 100% SL2. (*A*) All genes; (*B*) non-operon genes; (*C*) first genes in operons; (*D*) downstream genes in operons.

Association of long ICRs with long genes

Most ICRs are 50–200 nt long (median = 129) (Fig. 4A). However, several operons have very long ICRs, some even >2 kb. One possible reason why some ICRs are unusually long could be that they are present in expanded regions of the genome and are typically found closer to the ends of the chromosomes (Prachumwat et al. 2004). To test this idea, we looked for an association between long ICRs and long genes. Because introns are noncoding transcribed regions, like ICRs, intron

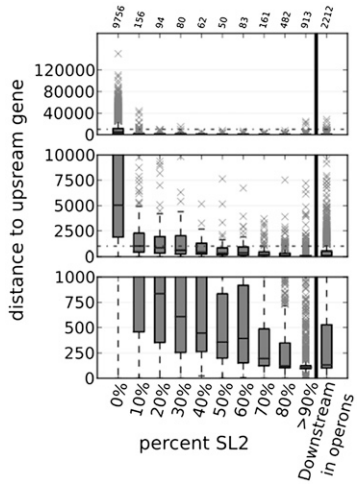

Figure 3. SL2 *trans*-splicing as a function of distance between genes. Distance between the polyadenylation site of the upstream gene and the *trans*-splice site of the downstream gene (*y*-axis) vs. the SL2% of the downstream gene. Box and whiskers plot in which the *top* of the box equals the first quartile and the *bottom* is the third quartile. The median value is denoted by a line *within* the box. The whiskers are 1.5× the inner quartile range. Outliers are represented by ×. All three panels show the same data with progressive magnifications of the lower ranges. The *top* panel shows all values. The numbers *above* the *top* panel are the numbers of genes in each category. The dotted line denotes the maximum value of the *middle* panel, in which only values below 10,000 are plotted. Similarly, the *bottom* shows only values below 1000. On the *right*, the data for all downstream operon genes are shown.

length was used as a proxy for gene length. The ICRs were divided into bins based on their length, and the number of ICRs in each bin is shown above the bar in Figure 4B. The boxes show a direct relationship between ICR length and intron length of the adjacent genes. This suggests that expanded ICRs are associated with expanded genes and, therefore, are likely to be a consequence of the same phenomenon. Although this finding clearly indicates that some ICRs are long because they are in regions of the genome that have been expanded, in the next section we describe analysis that indicates that sometimes ICRs are longer than usual to accommodate an internal promoter.

SL percentage and internal operon promoters

SL1 *trans*-splicing is believed to occur at the 3′ ends of outrons, adjacent to the promoter. In contrast, SL2 *trans*-splicing occurs downstream in operons, more than a gene's length from the promoter. Interestingly, however, some genes annotated as downstream in operons receive a mixture of SL1 and SL2. Could these be downstream in hybrid operons (Huang et al. 2007)? Perhaps pre-mRNAs receive SL2 when transcribed from promoters at the 5′ ends of operons, and SL1 when transcribed from a proximal promoter between the operon genes. This idea predicts a correlation between the presence of a proximal promoter and an SL1/SL2 mixture at a given *trans*-splice site. As a mark of promoters, HTZ ChIP data (Whittle et al. 2008) were used. HTZ is a histone H2A isoform that is present at some, but not all, promoter regions. In ChIP-chip experiments, Whittle et al. (2008) found 23% of annotated genes had an HTZ peak and 37% of operons contained internal peaks, possibly indicating internal promoters. *Trans*-splice sites were divided into bins based on the ratio of SL1/SL2 and the number of genes in each bin with an HTZ peak was plotted (Fig. 5A). The data clearly show that *trans*-splice sites with high SL1 are more likely to have HTZ peaks than those with high SL2. Furthermore, *trans*-splice sites with a mixture of SL's are the most likely to have an HTZ peak.

To explain this result, we suggest that genes that receive a significant percentage of SL1 are transcribed from a proximal promoter. However, why are genes that receive a mixture of SL1/SL2 more likely to be within an HTZ peak than are isolated genes that get high SL1? Could this be due to our requirement that the *trans*-splice site be within an HTZ peak? Since the HTZ peak can be anywhere upstream of the *trans*-splice site, limiting the promoter position to near the *trans*-splice site could affect our ability to detect promoters. For genes that receive a mixture of SL1/SL2 (in operons), the relatively short intergenic distance could make these genes not directly comparable to genes that receive primarily SL1. They can have a much longer distance between genes, allowing the promoter to be farther from the *trans*-splice site. To control for this possibility, the data set was limited to genes that had <500 nt distance between them. The percentage of genes with promoters detected goes up significantly when the distance between genes is limited (Fig. 5B). The percentage of genes with high SL1 in HTZ peaks is nearly double the percentage of all genes with HTZ peaks (Whittle et al. 2008), while the genes with mixed SL1/SL2 are still the most likely to be in an HTZ peak, and thus to have an internal promoter. Finally, the percent of SL2 genes that have an HTZ peak does not increase substantially. This analysis suggests that HTZ is even more likely to be associated with promoters within hybrid operons than it is with non-operon genes, a result for which we lack an explanation. Surprisingly, some genes with high SL2 have HTZ peaks over the *trans*-splice site. However, the HTZ peaks associated with these genes tend to be much larger and to have lower z-scores than the peaks for the genes that get higher SL1 (data not shown). Longer peaks may be large enough to be associated with promoters of other genes, while lower z-scores may indicate that these HTZ peaks do not mark actual promoters. In any case, these data make it clear that higher levels of SL1 *trans*-splicing are associated with the presence of proximal promoters.

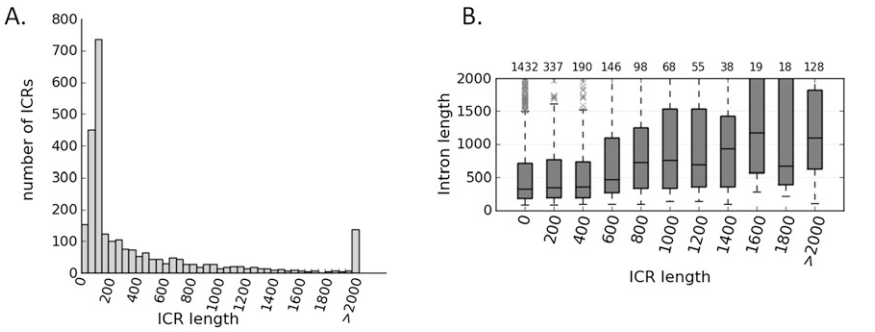

Figure 4. ICR length vs. gene length. (*A*) Histogram of all ICRs shows that most ICRs are <200 nt long. ICR length is in bins of 50 nt. Number of ICRs with the indicated length is shown on the *y*-axis. (*B*) Box-and-whiskers plot of average intron length of the genes upstream and downstream of the ICR vs. the ICR length. The number of genes in each category is indicated by numbers *above* the graph. ICR lengths are in bins of 200 nt.

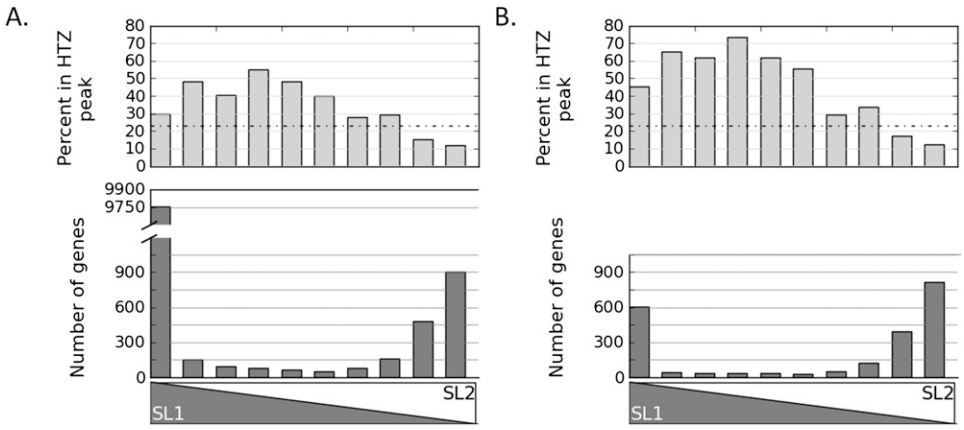

Figure 5. Relationship between *trans*-splicing specificity and proximal promoters. (*Bottom*) Histogram of the total number of genes with indicated SL1/SL2 proportion. (*Top*) The percent of those genes associated with a peak of the minor histone HTZ, which marks promoters (Whittle et al. 2008). (*A*) Genes with low percent SL2 (high SL1) are more likely to be in an HTZ peak than genes with high SL2. *Trans*-splices sites with mixed SL are the category most likely to be in an HTZ peak. (*B*) When genes with <500-bp intercistronic distance are analyzed, those with high SL1 are even more likely to be in an HTZ peak. Dotted line indicates the percent of all genes with an HTZ peak (Whittle et al. 2008).

A clear expectation of the presence of internal promoters would be that genes downstream of internal promoters would have increased expression compared with the gene just upstream. However, expression of operon genes drops somewhat going from the 5′ to the 3′ end of the cluster (Cutter et al. 2009). This could be due to inefficient processing of the polycistronic precursor or failure of transcription. Nonetheless, when we plot the difference between expression levels of operon gene pairs vs. the percent SL2 (Supplemental Fig. 6), it is clear that hybrid operons (those with higher SL1 *trans*-splicing) have a much smaller differential between the gene pairs. This indicates that the internal promoters boost expression of the genes downstream from them, as would be expected.

Deletions of promoters in hybrid operons change the SL1/SL2 ratio

It seems likely that genes with mixtures of SL1/SL2 have two promoters: the internal promoter, which creates the majority of transcripts *trans*-spliced to SL1 and the promoter at the 5′ end of the operon, which creates the transcripts *trans*-spliced to SL2. This idea was tested using mutations upstream of genes with SL1/SL2 mixtures with deletions in either the promoter at the 5′ end of the operon or the proposed internal promoter.

The deletion *sptl-1(ok1693)* is within an operon and deletes a 963-bp region ending just 40 nt upstream of the *trans*-splice site of *sptl-1*, which receives 71% SL1/29% SL2 in the deep-sequencing data. In addition, there is an HTZ peak associated with this ICR. Thus, we propose that *sptl-1* has an internal promoter that should be deleted by the ok1693 mutation, since more than two-thirds of the 1547-bp ICR upstream is deleted (Fig. 6A). The level of *sptl-1* RNA is, in fact, reduced approximately fourfold in the deletion strain, as determined by RT–PCR. Furthermore, the ratio of SL1 vs. SL2 is dramatically shifted by the deletion (Fig. 6A). SL1 is much lower in the mutant strain compared with SL2, consistent with the idea that the SL1 is *trans*-spliced to pre-mRNA from the now deleted internal promoter. However, there are two alternative explanations for this finding: the shortened ICR could promote SL2 *trans*-splicing or sequences favoring SL1 *trans*-splicing could have been removed by the deletion. Nonetheless, the most parsimonious explanation for the

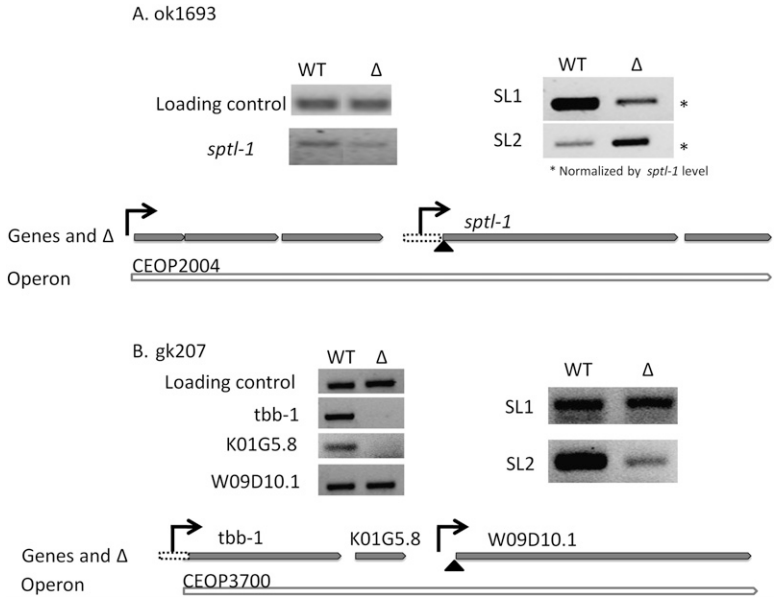

Figure 6. Effect of promoter deletions on the SL1/SL2 ratio. Diagrams at the *bottom* of each panel are to scale. They show pointed gray rectangles to indicate genes, hollow pointed empty rectangles to indicate operons and dotted boxes to indicate deletions. Black arrows indicate predicted locations of promoters and the triangle indicates the *trans*-splice site analyzed by RT-PCR. (*A*) RT-PCR in wild type (WT) and *ok1693* (Δ). To make visualization of SL1/SL2 ratio easier, *sptl-1* levels have been normalized to wild-type by adding 4× cDNA of the deletion strain based on the levels of mRNA for this gene (data not shown). (*B*) RT-PCR in wild type (WT) and *gk207* (Δ). (*Left*) RT-PCR of mature mRNAs. (*Right*) RT-PCR of W09D10.1 *trans*-spliced to SL1 and SL2. The same level of cDNA was used for all PCRs.

reduction in SL1 *trans*-splicing is deletion of the promoter in the ICR.

Would a deletion of the promoter at the 5′ end of an operon lower the level of SL2 *trans*-spliced product of a downstream gene that receives a mixture of SL1/SL2? The mutation *tbb-1(gk207)*, which deletes the majority of the region upstream of CEOP3700, was tested. This deletion also removes the *trans*-splice site of the first gene. The third gene in the operon, W09D10.1, has two *trans*-splice sites 6 nt apart, both of which receive mixtures of SL1/SL2. Based on RNA sequencing data, W09D10.1 receives 44% SL1/56% SL2. The mutation eliminates all expression of the first two genes in the operon, confirming the idea that the promoter has been deleted by this mutation (Fig. 6B). However, the third gene, which we predicted to have an internal promoter, does not lose expression. In addition, there is an HTZ peak associated with this ICR. Consistent with expectation, by removing the promoter at the 5′ end of the operon, the level of the SL2 *trans*-spliced products is dramatically reduced, whereas the SL1 *trans*-spliced products remain unchanged. These data are consistent with the idea that transcripts of downstream operon genes coming from the promoter at the 5′ end of the gene cluster are primarily SL2 *trans*-spliced, while transcripts coming from a proximal and internal promoter are essentially outron-containing transcripts and are, therefore, *trans*-spliced to SL1.

Rare *trans*-splicing to *cis*-splice sites

Although most *trans*-splice sites map to the 5′ ends of genes, some *trans*-splicing clearly also occurs at sites not associated with gene 5′ ends. This *trans*-splicing is relatively rare, and is therefore seen more easily with increased numbers of sequencing runs. Many of the less-abundant and newly recognized *trans*-splicing events mapped to annotated *cis*-splice sites, a phenomenon recognized previously both in vitro and in vivo (Choi and Newman 2006; Lasda et al. 2010). In fact, because *cis*- and *trans*- splice sites share a consensus sequence, it is somewhat surprising that *trans*-splicing at *cis*-splice sites is not even more prevalent.

In total, 11,157 *trans*-splicing events map to *cis*-splice sites (7% of all known *cis*-splice sites; Supplemental Table 1). At these sites, the level of *trans*-splicing is quite low as indicated by the fact that it can be seen easily when only a single read is required, but drops off dramatically when even only two reads are required (Supplemental Table 2). The vast majority of *C. elegans* introns are very small (47% of introns are between 41 and 60 bp) (Choi and Newman 2006). However, these rare *trans*-splicing events tended to occur in long introns, as previously reported by Choi and Newman (2006). All introns were grouped into bins based on their length, and the frequency of *trans*-splicing to sites in each bin was measured (Fig. 7). Clearly, very few of the small introns are ever *trans*-spliced. Furthermore, the larger an intron is, the greater the likelihood of a *trans*-splicing event.

Discussion

Analysis of the large collection of modENCODE RNA sequences has confirmed past observations about *trans*-splicing and has enhanced our understanding of the relationship between *trans*-splicing and other features of the genome including the distance between genes, the presence of promoters, and the annotation of operons. This global study of *C. elegans trans*-splicing confirms that SL1 and SL2 *trans*-splicing are distinct phenomena and that previous estimates of the percent of genes whose transcripts are subject to *trans*-splicing (*trans*-spliced genes) are accurate.

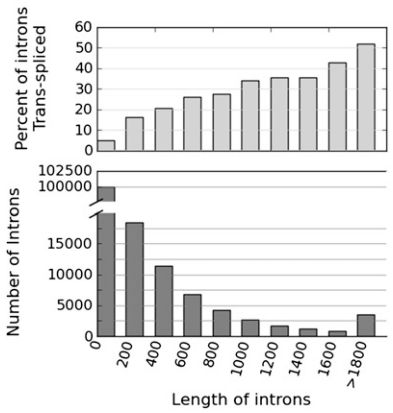

Figure 7. Rare *trans*-splicing at intron 3′ splice sites vs. intron length. Sites at which some *trans*-splicing occurred were mapped to annotated intron 3′ splice sites. Introns were divided into bins of 200 nt according to their length, with the number indicating the lower end of the bin. The *bottom* graph shows the number of introns in the genome with the indicated length, and the *top* graph shows the percent of the introns in each bin that have at least one associated *trans*-splicing event.

What percent of *C. elegans* genes are *trans*-spliced?

In 1994, it was estimated that 70% of *C. elegans* genes were *trans*-spliced (Zorio et al. 1994). We show that at least 56% of *C. elegans* genes are *trans*-spliced and we estimate, in total, ~70% are *trans*-spliced. However, there are two potential issues with our calculations. First, the lack of proper 5′ end annotations may affect the numbers. Second, detection of *trans*-splicing is dependent on sufficient expression. There is clearly a trend that using higher expression cutoffs results in a higher estimate of the percentage of genes *trans*-spliced. This trend could be due to higher levels of *trans*-splicing of highly expressed genes, or to the failure to detect *trans*-splicing of genes expressed at low levels. The former is most likely true because, when genes were binned by expression level, the higher the expression the more likely the gene was to be *trans*-spliced. When all *trans*-splice sites whose expression levels were >0.05 dcpm were included, the percentage of genes *trans*-spliced rose to 71%. Therefore, we predict ~70% of *C. elegans* genes are *trans*-spliced.

SL1 and SL2 *trans*-splicing are distinct phenomena

Either SL1 or SL2 dominates at the majority of *trans*-splice sites; a mixture is uncommon. The SL1/SL2 ratios clearly demonstrate that SL1 and SL2 *trans*-splicing are different phenomena. Both types of *trans*-splice sites may have surrounding sequences that signal SL1 or SL2 specificity. Alternatively, SL2 may be specified, while SL1 may be used as the default SL. We favor the latter idea because sites that receive a majority of SL1 are, in general, 100% *trans*-spliced to SL1, whereas sites that receive a majority of SL2 still receive significant levels of SL1. If SL1 is the default SL, the SL1 events at these preferentially SL2 sites could be due to the limited speed of SL2 *trans*-splicing or to imperfect specification of SL2. On the other hand, SL1 *trans*-splicing may occur more frequently at SL2-specified sites due to the ~10-fold higher levels of the SL1 RNA compared with the SL2 RNA (S Kuersten, R Conrad, T. Blumenthal, unpubl.). Finally, one other reason for suggesting SL1 is the default SL is that *cis*-splice sites, which are *trans*-spliced infrequently, tend to be *trans*-spliced to SL1 rather than SL2, although this could again be a consequence of the 10-fold higher level of the SL1 snRNP (data not shown).

Low-level *trans*-splicing at *cis*-splice sites

Cis-splice sites are *trans*-spliced very rarely, but more often when the intron is large. This *trans*-splicing could be due to weak promoters within the introns that cause a very low level of transcription (Choi and Newman 2006) and/or because the long introns are sometimes interpreted as outrons by the splicing machinery. Since any AU-rich sequence can act as an outron when inserted at the 5′ end of an mRNA (Conrad et al. 1995) and an outron can be converted to *cis*-splicing by inserting a 5′ splice site within it (Conrad et al. 1991), it is clear that *cis*- and *trans*-splice sites are interchangeable given the appropriate pre-mRNA context. We believe that long introns are more easily mistaken for outrons because of the increased physical distance between the 5′ and 3′ splice sites. If these *trans*-splicing events generally represent inaccuracies, as seems probable, this may explain why *C. elegans* introns are typically quite short (Blumenthal and Steward 1997). Introns in *C. elegans* may have shortened over evolutionary time to prevent inappropriate *trans*-splicing that would destroy transcripts by splicing within the transcript. In this regard, *trans*-splicing to *cis*-splice sites could be used to inactivate transcripts, depending on circumstances, resulting in a novel regulatory mechanism for genes with large introns. In addition, *trans*-splicing to *cis*-splice sites can also be used to create alternative isoforms of some mRNAs (Yin et al. 2010). However, the first AUG after the *trans*-spliced *cis*-splice sites is not in-frame more often than predicted, suggesting that this mechanism may not be used frequently (Supplemental File 1).

SL2 *trans*-splicing and the annotation of operons

A significant fraction of *C. elegans* genes are cotranscribed with other genes in operons. The most definitive way to demonstrate the existence of an operon would be to identify a polycistronic RNA. However, because of rapid processing of the pre-mRNA, both by 3′ end formation and *trans*-splicing between the genes, detecting polycistronic RNAs has proved possible only occasionally, so an alternative method must be used. So far, operons have been identified using the dual criteria of *trans*-splicing to SL2 and short intercistronic distance. This study demonstrates that annotation based on these criteria is, in general, quite accurate.

There is a very strong relationship between *trans*-splicing to high levels of SL2 and the presence of another gene nearby in the same 5′ to 3′ orientation at the whole-genome level. Furthermore, there is a strong relationship between the SL2% and the distance to the nearest upstream gene, an observation with significant implications for the mechanism of SL2 *trans*-splicing. In contrast, SL1-*trans*-spliced genes very rarely have closely spaced upstream genes in the same orientation. Clearly, most operons can be accurately identified using SL2 *trans*-splicing and close genes as criteria.

However, there are a few examples in the genome where SL2 *trans*-splicing occurs without a gene just upstream. These could be due to some operons having long spacing between genes. Long ICRs are in some cases the product of an expanded genic region. In these cases, the forces that either cause or allow introns to be long also seem to cause or allow ICRs to be long. Parenthetically, there are a few examples of SL2 *trans*-splicing to *cis*-splice sites at the 3′ end of the first intron, a rare phenomenon for which we do not currently have an explanation.

Genes *trans*-spliced to a mixture of SLI and SL2

The presence of high levels of SL1 is clearly linked with the presence of a proximal promoter, whereas the presence of high levels of SL2 is linked with the lack of a proximal promoter. Thus, SL2 *trans*-splicing is reserved for genes that are downstream in operons. Interestingly, however, there are many genes that receive a mixture of SL1 and SL2, and these genes are properly annotated as operons. These genes are members of hybrid operons (Huang et al. 2007). Genes with as low as 10% SL2 have a restricted distance to the upstream gene (Fig. 3). This implies that these genes are required to be close together and are therefore in operons. Also, genes that receive both SL1 and SL2 often have HTZ peaks associated with their *trans*-splice sites, indicating that there is a promoter between the genes.

The genes in these operons can most likely be transcribed from two different promoters. When a gene within a hybrid operon is transcribed from the promoter upstream of the operon, a polycistronic mRNA is created, and SL2 is *trans*-spliced. When the same gene is transcribed from the promoter located adjacent to it, the pre-mRNA has an outron and SL1 is *trans*-spliced. If the two promoters were used to similar extents, this would create the observed mixture of spliced leaders. This hypothesis is consistent with the results seen in the two deletion strains analyzed experimentally. When a presumed internal promoter was deleted, expression of the gene just downstream from it was dramatically reduced and the *trans*-splicing to SL1 even more dramatically reduced. The most likely explanation is deletion of the internal promoter, but we cannot eliminate the possibility that the fact that the genes were brought closer together is responsible for the reduction in SL1 *trans*-splicing. In contrast, when the promoter at the 5′ end of an operon was deleted, SL2 *trans*-splicing to the gene just downstream of the internal promoter was dramatically reduced. The small amount of SL2 *trans*-spliced mRNA remaining could come from the adjacent promoter or from residual promoter activity upstream.

Huang et al. (2007) concluded that *trans*-splicing with a mixture of SL1 and SL2 is not a sufficient indicator for the prediction of an internal promoter. This may well be true, but our data indicate clearly that such a mixture is associated with hybrid operons overall, and that it is therefore capable of predicting the presence of an internal promoter. Nonetheless, in some instances there may well be other reasons for a downstream operon gene to receive a mixture of the two spliced leaders. Additionally, genes with high levels of SL2 might also be within hybrid operons if the operon promoter were far stronger than the internal promoter.

In conclusion, analysis of RNA sequences indicates that around 70% of *C. elegans* genes are subject to *trans*-splicing. SL1 is *trans*-spliced to genes that are first genes in operons or not in operons at all, while SL2 is *trans*-spliced to genes that are downstream in operons. Furthermore, there is a link between the presence of an adjacent promoter and *trans*-splicing to SL1, and SL1/SL2 ratios can be used as indicators of hybrid operons. Within a hybrid operon, the use of SL2 is due to transcription originating at the 5′ end of the operon, whereas SL1 *trans*-splicing occurs on transcripts originating at the promoter internal to the operon. Consequently, the SL1/SL2 ratio should be one feature considered when predicting promoter locations of *trans*-spliced genes. To avoid certain experimental pitfalls, *C. elegans* researchers should consider whether their gene of interest is *trans*-spliced, and if so, the gene's position within an operon as well as the gene's SL1/SL2 ratio.

Methods

C. elegans strains

The following stages and strains of *C. elegans* were processed for RNA sequencing (RNA-seq): embryonic *him-8(e1489)* (50% males),

early embryos, late embryos, *lin-35(n1745)* L1, L1, L2, L3, dauer entry *daf-2(e1370)*, dauer *daf-2(e1370)*, dauer exit *daf-2(e1370)*, L4, L4 males, JK1107 L4 (no gonad) *glp-1(q224)*, young adults, aged adults [*spe-9(hc88)*], adults exposed to Harposporium spp (tentative assignment) (as well as a control exposed to *E. coli*), and adults exposed to *S. marcescens* (as well as a control exposed to *E. coli*) (Gerstein et al. 2010). All worms were N2 and grown on NGM plates and fed *E. coli* strain OP50, unless otherwise noted.

RNA isolation and transcriptome sequencing

RNA was extracted, reverse transcribed, and the resulting DNA was subject to deep sequencing on the Illumina platform as detailed in Hillier et al. (2009) and Gerstein et al. (2010). *Trans*-splice site locations, number of reads, and depth of coverage per million reads (dcpm) for each gene were determined as described in the Supplemental Methods of Hillier et al. (2009) and are outlined below.

Briefly, a database was created of all potential *trans*-splicing events. To create the database, the sequence of each SL was spliced to each possible acceptor in the genome (as predicted or annotated by GENEFINDER [run with permissive parameters], TWINSCAN, and WormBase). All RNA-seq reads were then matched to the database using cross_match, retaining only the database matches with score ≥24, ≤2 mismatches. The alignment to the spliced leader had to have a score of at least two better than any of the other hits to the SL database and at least five better than the alignments to the genome or splice junction database. At least one of the *trans*-spliced reads at each *trans*-splice site had to have nine bases matching the SL sequence. Number of reads refers to the number of sequencing reads matching the above criteria at a single site in the genome. Only 36-bp reads were used; no assembly was performed.

To calculate depth of coverage per million reads, dcpm, one first calculates depth of coverage, which is the approximate number of reads at a single position within the gene, analogous to fold coverage. This number is approximate because: (1) the number of reads across a gene is smoothed using overlapping windows, (2) both nonunique and below-threshold reads are excluded from this calculation (for details of calculation, see Supplemental Methods in Hillier et al. 2009). To make depth of coverage of a gene comparable across data sets, the depth of coverage is divided by the number of millions of aligned reads in the data set. One can recover the approximate fold coverage of a gene by multiplying the dcpm by the number of millions of aligned reads in the data set (generally 14–20 million). Average dcpm across all stages was used.

Programs used

The databases used for analysis of the *trans*-splice sites were built in MySQL Server version: 5.0.77 Source distribution. The calculations, manipulations, and the retrieval of the data were performed with Python 2.6.4 and IPython 0.8.4 using the python site-packages numpy (version 0.4.0rc1), cogent (version 0.4), and mysqldb (version 1.2.1) site packages. The graphs were drawn with either the python site-package matplotlib (version 0.99.1.1) or Microsoft Excel (version 12.0.6545.5000).

GO enrichment

GO profiling was used to identify statistically overrepresented and under-represented GO terms in the *trans*-spliced datasets. Briefly, two lists of genes with an average dcpm of >0.05 were created: a list of *trans*-spliced genes and a list of genes with no evidence of *trans*-splicing. The GOstat2 program (http://gostat.wehi.edu.au/) with the false discovery rate (Benjamini) correction was used to compare the genes in both lists, which had GO-terms (Beißbarth and Speed 2004). Level 3 and higher GO hierarchy were reported. Only GO terms with *P*-values of <0.05 are included in Supplemental File 4.

SL1/SL2/SL2 variant ternary plot

For each *trans*-splice site, the percentage of SL1, SL2, or SL2-variant reads at each position was calculated. The binomial exact test was used on the percent SL2 and samples with a *P*-value of >0.05 were removed. The percentage was plotted on a ternary plot that was based on code by written by C.P.H. Lewis while at the University of California, Berkeley, 2008–2009.

Ratio of SL1/SL2 (SL2+SL2 variant) and percent SL2 (SL2+SL2 variant)/(SL1+SL2+SL2 variant)

The number of SL1 and SL2 reads (including SL2 variants) for each site was totaled across stages to determine the number of reads for a given site. For each *trans*-splice site the ratio of SL1/SL2 was calculated by counting the number of SL1 reads divided by the number of SL2 reads. Sequences that corresponded with SL2 variants were included as SL2 reads in this calculation. The percent SL2, which refers to the number of SL2 reads divided by the number of total *trans*-spliced reads at a single position was then calculated. For Figures 2, 3, 5, and Supplemental Figure 4, a binomial exact test was used on each ratio and all of the sites with *P*-values more than 0.05 were removed.

Genomic units

The WormBase ws207 gff3 file was used to create lists of genes, operons, and introns (Supplemental Fig. 7). Positions of the genomic units were mapped to WormBase ws170 coordinates using unmap_gff_between_releases, (downloaded from http://www.sanger.ac.uk), which facilitated the conversion of the whole gff file to ws170 coordinates. The gene/transcript list was created by selecting all entries in the gff file in which the feature was protein_coding_primary_transcripts or pseudogene, and subselecting transcript name/gene name out of the line_group. A list of all operons in the gff file was created by selecting all entries in which the feature was operon and subselecting the operon name out of the line_group. The list of genes in operons was created by selecting all genes whose coordinates were completely within the coordinates of an operon and on the same strand as the operon. Gene order (position within each operon) was discovered by ordering the genes in each operon by their 5′ end. If two genes had the same 5′ end, the shorter of the two genes was used as the first gene, because often the longer gene was a misannotation of gene structure. If a single gene had multiple 5′ ends, the most 5′ of the 5′ ends was used. Finally, the table, trans_gene_ops, was created that contained every transcript/gene in the gene/transcript list, the location of the 5′ and 3′ ends of the transcript, the operon the transcript was contained within (if applicable), and the position of the gene within the operon (if applicable).

A 3′ *cis*-splice site list was created by collecting all entries in the gff file, where the feature was intron and the source was coding_transcript. The 3′ end positions of all introns were collected. To find the 3′ splice site, the position of the last nucleotide of the introns was corrected by 1 nt (for introns on the positive strand one was added and for those on the negative strand one was subtracted). All *cis*-splice sites that were also annotated as the 5′ end of a protein-coding primary transcript, or a pseudogene were removed. The introns were used to find the 3′ splice sites instead of

the exons, because if exons had been used, all first exons would have needed to be removed. Finally, a file that contained the frame of the *cis*-splice site was created by selecting all of the exons from the gff file in which the source was coding_transcript and the feature was exon and a frame was indicated.

Operon positions

First genes and downstream genes of operons were determined by first ascertaining the positions of all genes that mapped within previously annotated operons. To belong to an operon a gene had to be completely within the operon's coordinates and be on the same strand. All genes that did not map completely within an operon were considered non-operon genes. Within each operon, the order of the genes was determined by the position of the 5′ end of each gene in the operon. If a gene had multiple isoforms with multiple 5′ ends, the most 5′ position was used. If two genes' 5′ ends were at the same position in the database, the order of genes in the operon was manually curated. Genes with multiple *trans*-splices are included multiple times, so that each *trans*-splice site would be represented.

Categorizing of the *trans*-splice sites

The *trans*-splice sites were categorized as either a "gene *trans*-splice site," a "*trans*-spliced *cis*-splice site", or "other." Briefly, for gene *trans*-splice sites, the site or the first AUG downstream of the site mapped within 500 nt of the 5′ end of a known gene. For *trans*-splicing *cis*-splice sites, the site mapped within 10 nt of a *cis*-splice site. All other sites were categorized as "other" sites.

Specifically, a table (potential_cis) was created of *trans*-splice sites that had the potential to be categorized as *trans*-spliced *cis*-splice sites. If a *trans*-splice site was within 10 nt of a 3′ *cis*-splice site, it was added to the potential *trans*-spliced *cis*-splice site list (potential_cis).

Discovering which *trans*-splice sites mapped to the 5′ end of genes was complicated by current annotations. Generally, the annotated 5′ end of *trans*-spliced genes in WormBase is either the *trans*-splice site or the start codon. Therefore, the position of the first AUG after each *trans*-splice site was determined (See section Length of 5′ UTRs). A list of known gene 5′ ends was created, which included 5′ ends: (1) found in the trans_gene_ops table, or (2) found in the ws170_from_ws207 gff whose feature was five_prime_UTR. Using the known 5′ ends, four values were calculated for each *trans*-spliced position: the distance from the *trans*-splice site to the closest downstream 5′ end and upstream 5′ end and the distance from the first AUG to the closest downstream 5′ end and upstream 5′ end. Of the four distances calculated, whichever distance was smallest was considered to be the distance to the closest gene. If the distance to the closest gene was <10 nt from the *trans*-splice site, that site was automatically annotated as belonging to that gene (Supplemental Fig. 8). Therefore, if any isoform of a gene was *trans*-spliced, the gene was counted as a *trans*-spliced gene.

Next, if the distance to the closest gene was less than 500, the site was annotated as belonging to the closest gene as long as there was not a *cis* splice site within 10 nt of the *trans*-splice site (as determined by the *trans*-splice sites in the potential_cis list). In that case, it was annotated as a *trans*-spliced *cis*-splice site. Next, any *trans*-splice site not previously annotated as a *trans*-spliced *cis*-splice site was annotated as a *cis* site if it was present in the potential_cis list. All other sites were annotated as "other." The question tree in Supplemental Figure 8 shows how the sites were categorized.

Frequency of *trans*-splicing at each *trans*-splice site

The number of non-*trans*-spliced reads for each *trans*-splice site at the 5′ end of a gene was determined by counting the number of reads whose entire sequence mapped exactly to the genome and whose 5′ end was between −22 and −9 upstream of a *trans*-splice site and in the same orientation. These positions were used to ensure that the non-*trans*-spliced genomic reads would be in an equivalent position to the *trans*-spliced reads. Since the length of a spliced leader is 22 nt, and since at least a 9-nt overlap was required to assign a spliced leader, the positioning of the genomic reads allowed direct comparisons of abundance. Frequency of *trans*-splicing was calculated as the number of *trans*-spliced reads divided by the sum of the *trans*-spliced reads and the non-*trans*-spliced reads. If a gene had more than one *trans*-splice site, we annotated the site with the higher *trans*-splicing frequency.

Trans-splice sites within HTZ peaks

The HTZ peak data from Supplemental Table 3 of Whittle et al. (2008) was used as a list of HTZ peak coordinates. The positions were converted to WormBase ws170 coordinates using remap_gff_between_releases, which was downloaded from the Sanger Center website. A *trans*-splice site was marked as having an HTZ peak only if it mapped within the peak. Nearby peaks were not used because downstream genes in operons could be close to the peaks of the upstream gene, especially if the peak is large and the first gene is small.

Distance to the upstream gene

The 3′ ends of the genes were determined by the Bartel Lab (Jan et al. 2010) with poly(A)-position profiling by sequencing (3P-Seq). Briefly, biotinylated oligos were ligated onto the 3′ ends of RNAs. RNAs were then reverse transcribed with a primer to the ligated oligo and dTTP as the only nucleotide. RNase H was used to digest the reverse-transcribed products, thereby releasing only the poly(A)-tailed mRNAs from the biotin. Those released mRNAs were used as substrate for RNA-seq. Distance to the next upstream gene is the position of the most abundant polyadenylation site of the upstream gene to the position of the *trans*-splice site. In operons, this is also called the length of the ICR.

Intron length of genes around ICRs

The intron length for the genes adjacent to each ICR was calculated. Intron length, in this case, refers to the sum of (1) the average intron length of all of the introns in the upstream gene, and (2) the average intron length of all of the introns in the downstream gene. If either of the genes did not contain an intron, it was excluded from the intron length analysis.

RT-PCR to determine SL use in mutants

The *tbb-1(gk207)* and *sptl-1(ok1693)* strains were obtained from the *Caenorhabditis* Genetics Center. Strains were grown on NGM plates spread with OP50. Mixed stage worms were isolated by washing populations off plates. Worms were immediately frozen at −80° in 4× volume of TRIzol. The pellets were thawed, mixed, and refrozen in liquid nitrogen three times. After thawing one last time, RNA was isolated according to the TRIzol protocol. The RNA was treated with DNase, cleaned by phenol-chloroform extraction, and then reverse transcribed with random primers according to the SuperScript II protocol. A dilution series of the cDNA was used to confirm that PCR reactions were in the linear range. A PCR for the control gene *rpl-26* was first performed at several cycle numbers to confirm that PCR was in the linear range. Primers used are

given in Supplemental Table 3. The amount of cDNA used for PCR with SL1 and SL2 in the *sptll-1* gene PCRs was normalized according to the levels seen in the *rpl-26* RT–PCR. The *tbb-1(gk207)* PCRs were not normalized to *rpl-26*, as the level of W09D10.1 was not changed significantly in the deletion strain. At least three biological replicates for each strain were performed.

Length of 5′ UTRs

To determine the length of the 5′ UTRs of *trans*-spliced genes, the position of the first AUG after each *trans*-splice site was established. First, the position of each AUG in the ws170 genome was found by searching the FASTA-formatted ws170 file for ATG and its reverse complement. This list was then compared with the positions of the *trans*-splice sites to determine the position of the first AUG 3′ to the *trans*-splice site on the same strand. The sequence of the SL plus the sequence from the *trans*-splice site to the first AUG is considered the sequence of the 5′ UTR in this context.

Predicting operons

To create a computationally predicted list of operons, each annotated gene was analyzed. If a gene had multiple *trans*-splice sites, a cumulative SL2 percentage was calculated. If a gene was *trans*-spliced to >10% SL2, it was predicted to be within an operon with its upstream gene. If multiple genes in a row were *trans*-spliced to SL2, they were all included in the operon. The first gene of the computationally predicted operons had to be either not recognized as *trans*-spliced or *trans*-spliced to ≥90% SL1. The predicted operon list was then compared with the ws207 annotated operons. The operons of the two lists were divided into four categories: exact match operons, partial match operons, annotated operons that had no overlap with the predicted operons, and predicted operons that had no overlap with the annotated operons. The exact match operon list contains only operons that have the exact same set of genes in both operon lists. The partial match operon list contains operons that have at least one gene in an operon in both lists. However, partial matches were only counted in the summary if there were two overlapping genes (Supplemental Fig. 5; Supplemental File 3). The other two categories, (annotated operons that had no overlap with the predicted operons and predicted operons that had no overlap with the annotated operons) contain only genes that are not in the other operon list.

Acknowledgments

This work was supported by research grants R01 GM42432 from the National Institutes of General Medical Sciences to T.B. and 1U01HG004263-01 from the National Human Genome Research Institute (NHGRI) model organism ENCylopedia of DNA elements (modENCODE) project to R.H.W. Some nematode strains used in this work were provided by the *Caenorhabditis* Genetics Center, which is funded by the NIH National Center for Research Resources (NCRR). We thank members of the Blumenthal lab for helpful discussions and critical reading of the manuscript.

References

Beißbarth T, Speed TP. 2004. GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20:** 1464–1465.

Blumenthal T. 2005. Trans-splicing and operons. *WormBook* **25:** 1–9.

Blumenthal T, Spieth J. 1996. Gene structure and organization in *Caenorhabditis elegans*. *Curr Opin Genet Dev* **6:** 692–698.

Blumenthal T, Steward K. 1997. Trans-splicing. In C. elegans *II*, pp. 129–132. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. 2009. Unlocking the secrets of the genome. *Nature* **459:** 927–930.

Choi J, Newman AP. 2006. A two-promoter system of gene expression in *C. elegans*. *Dev Biol* **296:** 537–544.

Conrad R, Thomas J, Spieth J, Blumenthal T. 1991. Insertion of part of an intron into the 5′ untranslated region of a *Caenorhabditis elegans* gene converts it into a *trans*-spliced gene. *Mol Cell Biol* **11:** 1921–1926.

Conrad R, Lea K, Blumenthal T. 1995. SL1 *trans*-splicing specified by AU-rich synthetic RNA inserted at the 5′ end of *Caenorhabditis elegans* pre-mRNA. *RNA* **1:** 164–170.

Cui M, Allen MA, Larsen A, Macmorris M, Han M, Blumenthal T. 2008. Genes involved in pre-mRNA 3′-end formation and transcription termination revealed by a *lin-15* operon Muv suppressor screen. *Proc Natl Acad Sci* **105:** 16665–16670.

Cutter AD, Dey A, Murray RL. 2009. Evolution of the *Caenorhabditis elegans* Genome. *Mol Biol Evol* **26:** 1199–1234.

Gerstein MB, Lu ZJ, Von Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE Project. *Science* **330:** 1775–1787.

Guang S, Bochner AF, Burkhart KB, Burton N, Pavelec DM, Kennedy S. 2010. Small regulatory RNAs inhibit RNA polymerase II during the elongation phase of transcription. *Nature* **465:** 1097–1101.

Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH. 2009. Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*. *Genome Res* **19:** 657–666.

Huang P, Pleasance ED, Maydan JS, Hunt-Newbury R, O'Neil NJ, Mah A, Baillie DL, Marra MA, Moerman DG, Jones SJ. 2007. Identification and analysis of internal promoters in *Caenorhabditis elegans* operons. *Genome Res* **17:** 1478–1485.

Jan CH, Friedman RC, Ruby JG, Bartel DP. 2010. Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. *Nature*. doi: 10.1038/nature09616.

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

Lasda EL, Allen MA, Blumenthal T. 2010. Polycistronic pre-mRNA processing in vitro: snRNP and pre-mRNA role reversal in *trans*-splicing. *Genes Dev* **24:** 1645–1658.

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

Prachumwat A, DeVincentis L, Palopoli MF. 2004. Intron size correlates positively with recombination rate in *Caenorhabditis elegans*. *Genetics* **166:** 1585–1590.

Spieth J, Brooke G, Kuersten S, Lea K, 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.

Sutton RE, Boothroyd JC. 1986. Evidence for trans splicing in trypanosomes. *Cell* **47:** 527–535.

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

Whittle CM, McClinic KN, Ercan S, Zhang X, Green RD, Kelly WG, Lieb JD. 2008. The genomic distribution and function of histone variant HTZ-1 during *C. elegans* embryogenesis. *PLoS Genet* **4:** e1000187. doi: 10.1371/journal.pgen.1000187.

Williams C, Xu L, Blumenthal T. 1999. SL1 trans splicing and 3′-end formation in a novel class of *Caenorhabditis elegans* operon. *Mol Cell Biol* **19:** 376–383.

Yin J, Yu L, Savage-Dunn C. 2010. Alternative *trans*-splicing of *Caenorhabditis elegans* sma-9/schnurri generates a short transcript that provides tissue-specific function in BMP signaling. *BMC Mol Biol* **11:** 46. doi: 10.1186/1471-2199-11-46.

Zorio DA, Cheng NN, Blumenthal T, Spieth J. 1994. Operons as a common form of chromosomal organization in *C. elegans*. *Nature* **372:** 270–272.

Received August 10, 2010; accepted in revised form November 19, 2010.