

# *Caenorhabditis elegans* Operons Contain a Higher Proportion of Genes with Multiple Transcripts and Use 3' Splice Sites Differentially

Fei Wang, Shi Huang, Long Ma\*

State Key Laboratory of Medical Genetics, Central South University, Changsha, China

## Abstract

RNA splicing generates multiple transcript isoforms from a single gene and enhances the complexity of eukaryotic gene expression. In some eukaryotes, operon exists as an ancient regulatory mechanism of gene expression that requires strict positional and regulatory relationships among its genes. It remains unknown whether operonic genes generate transcript isoforms in a similar manner as non-operonic genes do, the expression of which is less likely limited by their positions and relationships with surrounding genes. We analyzed the number of transcript isoforms of *Caenorhabditis elegans* operonic genes and found that *C. elegans* operons contain a much higher proportion of genes with multiple transcript isoforms than non-operonic genes do. For genes that express multiple transcript isoforms, there is no apparent difference between the number of isoforms in operonic and non-operonic genes. *C. elegans* operonic genes also have a different preference of the 20 most common 3' splice sites compared to non-operonic genes. Our analyses suggest that *C. elegans* operons enhance expression complexity by increasing the proportion of genes that express multiple transcript isoforms and maintain splicing efficiency by differential use of common 3' splice sites.

**Citation:** Wang F, Huang S, Ma L (2010) *Caenorhabditis elegans* Operons Contain a Higher Proportion of Genes with Multiple Transcripts and Use 3' Splice Sites Differentially. PLoS ONE 5(8): e12456. doi:10.1371/journal.pone.0012456

**Editor:** Ian Dunham, European Bioinformatics Institute (EBI), United Kingdom

**Received:** May 30, 2010; **Accepted:** August 5, 2010; **Published:** August 27, 2010

**Copyright:** © 2010 Wang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This study was funded by the National Science Foundation of China grant 30971639 and a research grant provided by the Central South University of China. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: longmace@gmail.com

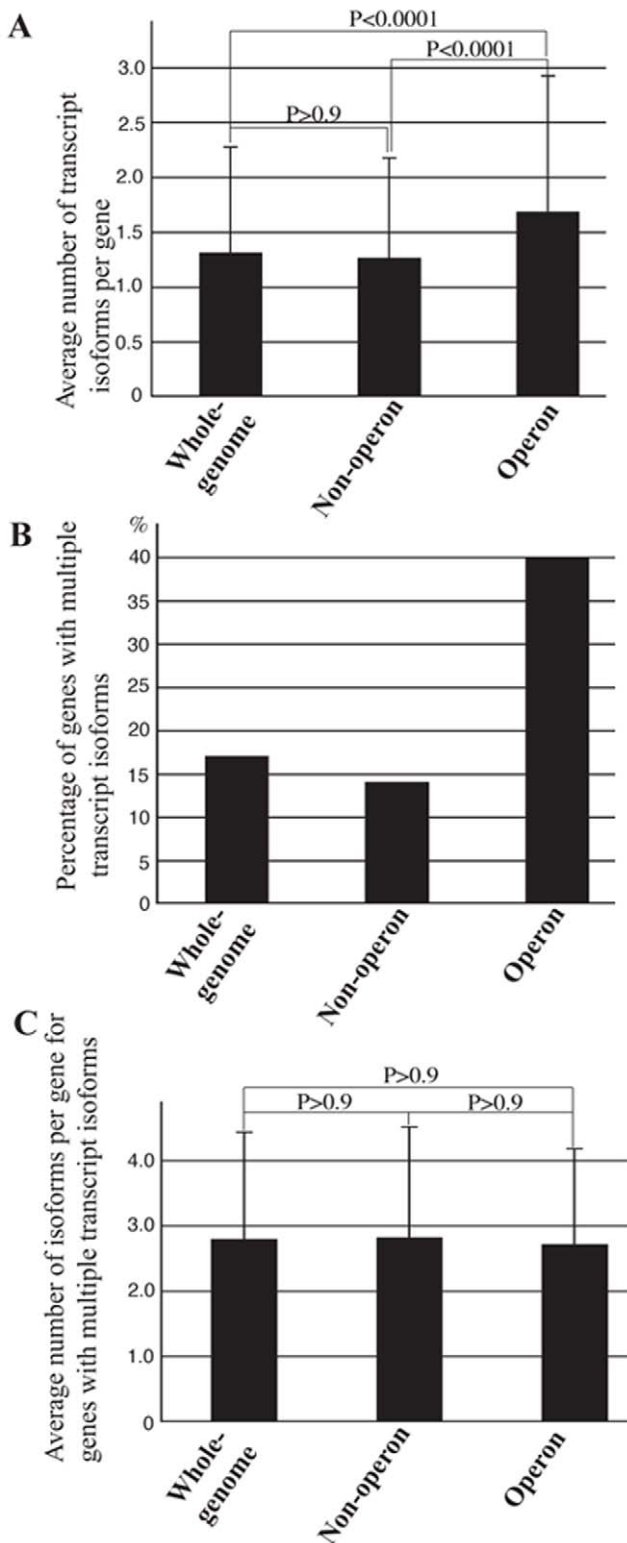
## Introduction

RNA splicing generates multiple transcript isoforms from a single gene and is believed to be a driving force for biological complexity in evolution [1,2]. In *C. elegans*, over 13% of genes are alternatively spliced [3]. In human, most genes are alternatively spliced [4,5,6]. Compared to RNA splicing, operons provide a different regulatory form of gene expression. An operon is a cluster of genes that are transcribed from a single promoter and controlled by the same regulatory sequences [7]. Operons exist abundantly in prokaryotes and are also found in eukaryotes, which include the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster* and some mammals [7,8]. In *C. elegans*, it was initially estimated that there were 15% of genes in about 1000 operons with an average of 2.8 genes per operon [9,10]. Recently the number of annotated operons in the *C. elegans* genome has increased to approximately 1250 (Wormbase Release 205), which gives an average of 2.3 genes per operon considering the number of operonic genes remains largely unchanged (around 2880, see the Results). In *C. elegans*, genes in an operon form a closely-spaced cluster with an ~100 bp intergenic distance [10]. However it is not known how operonic genes increase expression complexity, e.g., by RNA splicing, to adjust to the pressure of evolution and at the same time maintain their positional and regulatory relationships. *C. elegans* has a large number of operonic genes that are alternatively spliced, which provides an interesting model to understand the relationship between operons and RNA splicing.

## Results

We examined the average number of transcript isoforms per gene for genes of the whole genome, for all non-operonic genes and for all operonic genes. As shown in Figure 1A, non-operonic genes had about 1.26 transcript isoforms per gene, which was similar to the average of 1.31 transcript isoforms per gene for the whole genome. Operonic genes had 1.68 transcript isoforms per gene, which was over 30% more than that of the non-operonic genes.

One reason that operonic genes have more transcript isoforms per gene than non-operonic genes do is that operons may contain a higher proportion of genes that generate multiple transcript isoforms. Indeed, about 40% of all operonic genes have multiple transcript isoforms (Figure 1B and Table 1). However, only 14% and 17% of non-operonic genes and all genes, respectively, have multiple transcript isoforms (Figure 1B and Table 1). We next examined whether there is any difference in the average number of isoforms for genes that have multiple transcript isoforms. For all such non-operonic genes, there were about 2.81 isoforms per gene. For all such operonic genes, there were 2.71 isoforms (Figure 1C). For all genes of the whole genome, this number was 2.78, which was similar to that of operonic and non-operonic genes (Figure 1C). These results suggest that alternatively spliced operonic and non-operonic genes do not differ apparently in generating transcript isoforms. Therefore, operonic genes may utilize the splicing machinery as efficiently as non-operonic genes do to enhance their expression complexity.



**Figure 1. *C. elegans* operons contain a higher proportion of genes that express multiple transcript isoforms.** (A) *C. elegans* operonic genes express more transcript isoforms per gene than non-operonic genes do. (B) *C. elegans* operons contain a higher proportion of genes that express multiple transcript isoforms than non-operonic genes do. (C) Alternatively spliced *C. elegans* operonic genes and non-operonic genes have a similar number of transcript isoforms per gene. Z-test was performed (Figure 1A and 1C) to evaluate the significance of difference between the means of transcript numbers. Error bars represent standard deviations. doi:10.1371/journal.pone.0012456.g001

To investigate whether operonic introns utilize 3' splice sites differently from non-operonic introns, we analyzed the nucleotide sequences of position  $-7$  to  $-1$  of *C. elegans* introns. This sequence (3' splice site) is recognized by the splicing factors U2AF large and small subunits and plays important roles in regulating splicing efficiency and alternative splicing [11,12,13,14]. Among all 3' splice sites, the top 20 most commonly used sites were found in over 80% of introns (Table 2), suggesting that these sites are responsible for the splicing of the majority of introns. As shown in Figure 2, operonic introns use tttcag, attcag, ttccag and ttgcag significantly more frequently than non-operonic introns do, in which the frequency of ttgcag usage in operonic introns increased over 30% compared to that in non-operonic introns. 16 sites were used equally or less frequently in operonic introns. Among them, the frequencies of ttttag, gttcag, cttcag, atttag and gtccag were significantly reduced compared to that of non-operonic introns.

## Discussion

It is a challenge for operonic genes to increase expression complexity and maintain splicing efficiency while keeping strict positional and regulatory relationships. *C. elegans* operons may achieve these goals by at least two approaches. First, *C. elegans* operons significantly increase the proportion of genes that express multiple transcript isoforms (Figure 1). However, for genes that express multiple transcript isoforms, there is no apparent difference between the number of isoforms in operonic and non-operonic genes. This result suggests that *C. elegans* operons are more permissive for their genes to increase expression complexity by RNA processing than non-operonic genes are. By increasing the proportion of genes that express multiple transcript isoforms, *C. elegans* operons may compensate for a more strict transcriptional regulation and achieve the goal of expression complexity. Alternatively, *C. elegans* operonic genes may be under more pressure evolutionarily to enhance their transcript complexity, e.g., in order to perform more complex biological functions. Second, *C. elegans* operonic genes use four of the 20 most abundant 3' splice sites (tttcag, attcag, ttccag and ttgcag) more frequently and use the other 3' splice sites equally or less frequently (Figure 2). The differential usage of common 3' splice sites may help maintain efficient splicing of operonic genes, which are often highly expressed and have essential biological functions [9,10]. The differential usage of common 3' splice sites by operonic genes is also consistent with the notion that transcription and RNA splicing are coupled processes [1,2]. Compared to individual genes, it is plausible that the coupling of transcription and splicing of multiple genes in an operon presents a more challenging task for the splicing machinery, which may favor those 3' splice sites that optimize the splicing process and result in a differential use of common 3' splice sites by operonic genes.

The expression of transcript isoforms by *C. elegans* operonic genes may also depend on other regulatory mechanisms, e.g., by using different splicing silencers or enhancers and by generating alternative 5' and 3' untranslated regions (UTRs). Further analysis of these possibilities will provide a more comprehensive picture about the expression complexity of *C. elegans* operonic genes.

## Methods

We downloaded *C. elegans* gene names and annotated transcripts from the WormMart (WormBase Release 195) as html files. The data were processed using MS Excel to identify genes with different number of transcripts. Non-operonic genes were identified by deducting operonic genes from all genes of the whole genome. A random examination of over 100 operonic

**Table 1.** The numbers of genes and transcripts we analyzed.

	Whole-genome		Non-operon		Operon	
	Genes	Transcripts	Genes	Transcripts	Genes	Transcripts
<b>Genes with single transcript</b>	20109	20109	18369	18369	1740	1740
<b>Genes with multiple transcripts</b>	4248	11832	3106	8732	1142	3100
<b>Total</b>	24357	31941	21475	27101	2882	4840

Genes and annotated transcripts were downloaded from WormMart and processed with MS Excel.  
doi:10.1371/journal.pone.0012456.t001

genes that are annotated to have multiple transcript isoforms indicates that the isoforms for each gene share at least one coding exon.

The total number of each analyzed 3' splice site (positions -7 to -1) for the whole genome was obtained from the Intronerator (<http://genome-test.cse.ucsc.edu/Intronerator/>) [15]. We downloaded 16,087 unique operonic intron sequences from WormMart (WormBase Release 195) and processed the sequences using a software written in the C programming language and Microsoft Excel. Identical 3' splice sites (positions -7 to -1) are grouped and the proportion of each site is determined. The number of each 3' splice site for non-operonic genes was obtained by deducting the

number of the same site for operonic genes from the number for the whole genome. The online calculator for pairwise Z-test analysis is found at <http://www.dimensionresearch.com/resources/calculators/ztest.html>.

### Acknowledgments

We thank Linfeng Xia for processing *C. elegans* intron sequences.

### Author Contributions

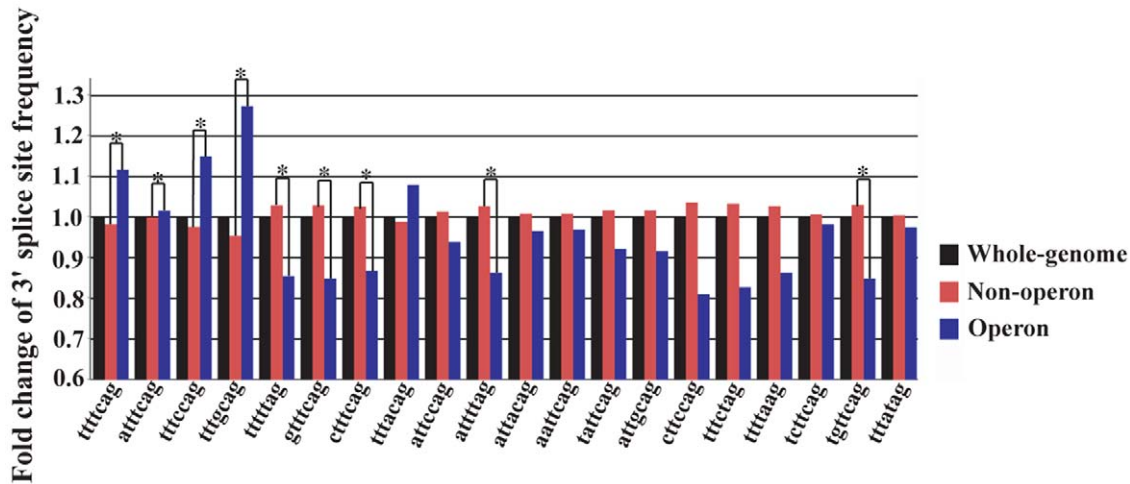
Analyzed the data: FW LM. Contributed reagents/materials/analysis tools: FW. Wrote the paper: SH LM.

**Table 2.** The proportions and numbers of the 20 most frequently used 3' splice sites in different groups of genes.

3' splice sites	Proportions and numbers of the top 20 3' splice sites in each group of genes						
	Whole-genome		Non-operon		Operon		<i>p</i> -value
	Ratio	n	Ratio	n	Ratio	n	
ttttcag	0.261	27890	0.256	23202	0.291	4688	<b>0</b>
atttcag	0.14	14919	0.139	12639	0.141	2280	<b>0.001</b>
tttccag	0.0857	9149	0.0835	7567	0.0983	1582	<b>0</b>
tttgacg	0.0359	3827	0.0341	3094	0.0456	733	<b>0</b>
tttttag	0.0346	3695	0.0355	3220	0.0295	475	<b>0</b>
gtttcag	0.034	3617	0.0348	3155	0.0287	462	<b>0</b>
ctttcag	0.0329	3508	0.0337	3050	0.0285	458	<b>0.001</b>
tttacag	0.0295	3152	0.0291	2640	0.0318	512	0.18
attccag	0.0295	3147	0.0298	2703	0.0276	444	0.18
attttag	0.0218	2327	0.0223	2025	0.0188	302	<b>0.002</b>
attacag	0.0161	1721	0.0162	1471	0.0155	250	0.37
aattcag	0.0153	1633	0.0154	1395	0.0148	238	0.36
tattcag	0.0137	1457	0.0138	1255	0.0126	202	0.33
attgcag	0.0129	1378	0.0131	1188	0.0118	190	0.04
cttccag	0.0115	1230	0.0119	1080	0.00932	150	0.03
tttctag	0.00949	1012	0.00978	886	0.00783	126	0.02
ttttaag	0.00866	924	0.00887	804	0.00746	120	0.23
tcttcag	0.00838	894	0.00841	762	0.00821	132	0.96
tggtcag	0.008	854	0.00822	745	0.00678	109	<b>0.01</b>
tttatag	0.00735	783	0.00737	668	0.00715	115	0.96
<b>Total</b>	<b>0.81628</b>	<b>87117</b>	<b>0.81085</b>	<b>73549</b>	<b>0.84225</b>	<b>13568</b>	NA

Total numbers of each 3' splice sites were calculated as described in Methods. The proportions of the top 20 sites were presented as a percentage of all identified 3' splice sites in the groups of genes specified. Pairwise Z-test was performed for each 3' splice site to test the significance of difference between proportions of operonic and non-operonic genes.  $p \leq 0.01$  is the confidence level.

doi:10.1371/journal.pone.0012456.t002



**Figure 2. Common 3' splice sites are used differentially by *C. elegans* operonic genes.** The proportions of each 3' splice site (X axis) of operonic and non-operonic genes were compared to that of all genes of the whole genome and were presented as fold changes (Y axis). Pairwise Z-test was performed (see Table 2) to evaluate the significance of difference between the proportions of each 3' splice site in operonic genes and non-operonic genes. \*:  $p \leq 0.01$ .  
doi:10.1371/journal.pone.0012456.g002

## References

1. Graveley BR (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* 17: 100–107.
2. Maniatis T, Tasic B (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418: 236–243.
3. Zahler AM (2005) Alternative splicing in *C. elegans*. *WormBook*.
4. Johnson JM, Castle J, Garrett-Engel P, Kan Z, Loerch PM, et al. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302: 2141–2144.
5. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40: 1413–1415.
6. Wang ET, Sandberg R, Luo S, Khrebukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456: 470–476.
7. Blumenthal T (1998) Gene clusters and polycistronic transcription in eukaryotes. *Bioessays* 20: 480–487.
8. Blumenthal T (2004) Operons in eukaryotes. *Brief Funct Genomic Proteomic* 3: 199–211.
9. Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, et al. (2002) A global analysis of *Caenorhabditis elegans* operons. *Nature* 417: 851–854.
10. Blumenthal T, Gleason KS (2003) *Caenorhabditis elegans* operons: form and function. *Nat Rev Genet* 4: 112–120.
11. Hollins C, Zorio DA, MacMorris M, Blumenthal T (2005) U2AF binding selects for the high conservation of the *C. elegans* 3' splice site. *RNA* 11: 248–253.
12. Kent WJ, Zahler AM (2000) Conservation, regulation, synteny, and introns in a large-scale *C. briggsae-C. elegans* genomic alignment. *Genome Res* 10: 1115–1125.
13. Ma L, Horvitz HR (2009) Mutations in the *Caenorhabditis elegans* U2AF large subunit UAF-1 alter the choice of a 3' splice site *in vivo*. *PLoS Genet* 5: e1000708.
14. Zhang H, Blumenthal T (1996) Functional analysis of an intron 3' splice site in *Caenorhabditis elegans*. *RNA* 2: 380–388.
15. Kent WJ, Zahler AM (2000) The intronerator: exploring introns and alternative splicing in *Caenorhabditis elegans*. *Nucleic Acids Res* 28: 91–93.