DOI: 10.1534/genetics.108.099283

Germline Expression Influences Operon Organization in the *Caenorhabditis elegans* Genome

Valerie Reinke*,[1] and Asher D. Cutter†

**Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520 and †Department of Ecology and Evolutionary Biology and the Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario M5S 3B2, Canada*

Manuscript received December 1, 2008
Accepted for publication January 26, 2009

ABSTRACT

Operons are found across multiple kingdoms and phyla, from prokaryotes to chordates. In the nematode *Caenorhabditis elegans*, the genome contains >1000 operons that compose ~15% of the protein-coding genes. However, determination of the force(s) promoting the origin and maintenance of operons in *C. elegans* has proved elusive. Compared to bacterial operons, genes within a *C. elegans* operon often show poor coexpression and only sometimes encode proteins with related functions. Using analysis of microarray and large-scale *in situ* hybridization data, we demonstrate that almost all operon-encoded genes are expressed in germline tissue. However, genes expressed during spermatogenesis are excluded from operons. Operons group together along chromosomes in local clusters that also contain monocistronic germline-expressed genes. Additionally, germline expression of genes in operons is largely independent of the molecular function of the encoded proteins. These analyses demonstrate that mechanisms governing germline gene expression influence operon origination and/or maintenance. Thus, gene expression in a specific tissue can have profound effects on the evolution of genome organization.

GENOME sequencing projects of multiple organisms provide the means to gain higher-order views of genome organization. As a case-in-point, recent work has demonstrated clearly that genes are not arranged randomly within genomes. Global gene expression studies in yeast, *Caenorhabditis elegans*, Drosophila, Arabidopsis, and humans concur that genes within a genomic neighborhood often exhibit similar expression profiles (Michalak 2008). For instance, sex chromosomes often have a paucity or enrichment of genes whose expression is regulated within the germline or by sexual identity (Reinke *et al.* 2000; Wang *et al.* 2001; Parisi *et al.* 2003). Additionally, local chromosome domains, ranging in size from 10 to several hundred kilobases, frequently contain genes that are coexpressed in specific tissues or under specific conditions (Cohen *et al.* 2000; Caron *et al.* 2001; Roy *et al.* 2002; Spellman and Rubin 2002). An open question from these studies is whether these expression patterns arise passively, such that neighboring genes adopt similar expression states as a by-product of local chromatin conformation, or whether natural selection drives genome organization by favoring coregulation of physically clustered genes.

In addition to the potential for chromatin states and selection for coregulation to induce coexpression, operons result in correlated expression of neighboring genes in multiple species, including flatworms, chordates, and nematodes such as *C. elegans* (Lercher *et al.* 2003; Michalak 2008). A genome-scale study determined the composition of >1000 *C. elegans* operons, which together contain ~15% of all genes (Blumenthal *et al.* 2002). In these operons, two to eight genes share a promoter and are transcribed as one pre-mRNA, which is rapidly processed into multiple single-gene transcripts in the nucleus prior to translation (Spieth *et al.* 1993; Zorio *et al.* 1994). Operons are evolutionarily stable once formed, with 96% of *C. elegans* operon structures being conserved in *C. briggsae*, either due to the difficulty of operon genes regaining proper regulation on their own or because the dissociation of their component genes into autonomous units is deleterious to the organism (Stein *et al.* 2003).

Bacterial operons often contain genes whose protein products function in a common process; for instance, genes in the *lac* operon are all required for lactose metabolism. By contrast, genes within an individual *C. elegans* operon commonly do not have obviously related functions. Although genes in *C. elegans* operons frequently encode proteins required for basic cellular processes—such as metabolism, transcription, and RNA processing—many genes functioning in these same processes are not in operons (Blumenthal and Gleason 2003). Additionally, transcripts from genes within a *C. elegans* operon show poor coexpression relative to bacterial operons (Lercher *et al.* 2003). Poor coexpression probably occurs in part because *trans*-splicing rapidly

[1]*Corresponding author:* Department of Genetics, Yale University School of Medicine, 333 Cedar St., New Haven, CT 06520.
E-mail: valerie.reinke@yale.edu

cleaves the polycistronic transcript into monocistronic messages, allowing independent post-transcriptional regulation, and in part because downstream genes in operons can have internal promoters (Huang *et al.* 2007; Whittle *et al.* 2008). Thus, an understanding of why and how *C. elegans* operons are organized has remained elusive, as no common functional characteristics of operons have been identified.

This report analyzes global gene expression data to demonstrate that one common property of almost all *C. elegans* operons is expression in the germline. The distal end of the hermaphrodite gonad contains undifferentiated germ cells that are capable of producing either sperm or oocytes. In the proximal gonad of fourth stage (L4) larvae, germ cells differentiate into spermatocytes over a brief (~6-hr) period. Once the animal reaches adulthood, germ cells entering the proximal gonad continuously differentiate into oocytes, as long as sperm are present for fertilization (~3 days). Previous microarray studies defined genes with two principal profiles of germline-enriched expression (Reinke *et al.* 2000, 2004). The first profile is characterized by expression during the fourth larval stage in germ cells undergoing spermatogenesis. The second profile—here generally termed "germline"—is characterized by expression in any nonspermatogenic germ cell, including undifferentiated germ cells in the distal gonad and developing oocytes in the proximal gonad.

The data presented here illustrate the propensity for operons to contain genes expressed in the germline, except for spermatogenesis genes, which almost never reside in operons. Additionally, operons tend to cluster along chromosomes and are frequently found clustered with monocistronic germline-enriched genes. Inclusion of genes in operons is more tightly associated with germline expression than with the molecular function of the encoded proteins, suggesting that the evolution of operon gene composition is driven more by germline expression than by protein function.

MATERIALS AND METHODS

Microarrays: The microarray data used in this analysis are described in detail in Reinke *et al.* (2004). Briefly, RNA from wild-type and *glp-4(bn2)* mutant hermaphrodites was isolated at various developmental stages and compared using DNA microarrays containing 94% of predicted *C. elegans* genes. Because *glp-4(bn2)* hermaphrodites lack germ cells, genes with increased expression in wild type relative to *glp-4(bn2)* are considered germline enriched. Additional microarray comparisons between hermaphrodites producing only sperm [*fem-3(gf)*] and hermaphrodites producing only oocytes [*fem-1(lf)*] identified genes with spermatogenesis- and oogenesis-enriched expression, respectively. The germline gene set includes all wild-type/*glp-4*- and all *fem-1*/*fem-3*-enriched genes, specifically excluding the *fem-3*/*fem-1*-enriched gene set (spermatogenesis genes). The numbers of genes in each set presented in this report are based on the updated annotations in WormBase (WS190).

Expression calculations for operons: WormBase version WS190 was used for all analyses. The representation factor in Figure 1A was calculated by dividing the observed number of overlapping genes from two independent groups by the number expected to overlap. The significance was determined by a hypergeometric probability distribution test, which determines the probability that the overlap between two data sets occurs randomly (the web application used in this study, along with the description and formulation of this test, can be found at http://elegans.uky.edu/MA/progs/overlap_stats.html). To determine the number of germline-expressed operons, a combination of microarray (Reinke *et al.* 2004) and *in situ* hybridization (http://nematode.lab.nig.ac.jp/dbest/srchbyclone.html) data was used (Figure 1B; images can be viewed by entering the name of the EST listed for each gene in supplemental data file 1). Genes lacking microarray and/or *in situ* data were not included in the percentages calculated in the text. Missing *in situ* data are due to the fact that either no assay was performed or no staining was detectable. No detectable staining can be attributed to either experimental failure or low sensitivity (the gene is expressed at low levels or in very few cells). If the 185 genes with no visible staining are dismissed as true negatives (experimental failure) and removed from the calculation, 95% of operon genes would be inferred to have germline expression (1887 germline operon genes of a total 1988 examined operon genes). Alternatively, all 185 genes could be expressed solely in somatic tissues but fail to be detected due to sensitivity. Including the 185 "no staining" genes as "somatic" sets the lower bound of the calculation (1887/2173) to 87%.

Functional classes: The complete list of genes in each category is included in supplemental data file 2. The chosen functional categories were based on Gene Ontology (GO) annotation using both molecular function and subcellular localization categories to avoid biases. Each category included at least a few germline-eniched genes. For this analysis, genes with spermatogenesis-enriched expression are grouped with the somatic category. For each functional category, the microarray and *in situ* hybridization data were taken into account as described above.

Clustering statistics: Operon clustering analyses were conducted by partitioning each chromosome into 100-kb-long nonoverlapping windows. The numbers of operonic (k) and monocistronic (m) protein-coding genes in each window, inferred from WormBase release WS190, were then used to calculate the binomial probability (P) of observing k or more operon genes in the interval given their overall frequency on the chromosome (p_c) and the total number of genes ($n = m + k$). Alternatively, the genomic operon frequency (p_g) was used. P was computed with the cumulative binomial density function $P[x \geq k] = \Sigma \binom{n}{k} p^k (1-p)^{n-k}$ with a Perl script. Significance of a given interval was evaluated with a false discovery rate of 0.05 for the resulting distribution of P-values, using the program QVALUE (http://genomics.princeton.edu/storeylab/qvalue/) (Storey and Tibshirani 2003). Analyses with alternative window sizes yielded qualitatively identical results.

The physical correspondence of operons with monocistronic "germline genes," as inferred from DNA microarrays (Reinke *et al.* 2004), was performed by first calculating the observed distance to the nearest operon for each germline-expressed monocistron. The germline-expression status of each monocistronic gene was then randomly permuted 1000 times, with the distance between germline monocistrons and nearest operons being recalculated for each permutation. Each chromosome was permuted separately to account for differences among chromosomes in the abundance of germline genes, and only protein-coding genes were considered. We then compared the distribution of mean distances for the

permuted data to the observed distances to test for an effect of germline-expression status of monocistronic genes on proximity to operons, controlling for gene density. We also calculated the fraction of monocistronic genes that had germline expression (f_g) for each of the 100-kb regions described above. Likewise, the number of operons divided by the total gene complement (f_o) of the interval was computed, which we used in nonparametric correlation with f_g, assuming independence of the nonoverlapping windows.

Analysis of 5′ flanking sequence: Sequence upstream of monocistronic genes and the first gene in operons was used in this analysis. Genes contained entirely within the intron(s) of larger genes were excluded from the analysis, as were noncoding RNA genes. We calculated the lengths of the 5′ intergenic regions with a customized Perl script as a proxy for upstream regulatory complexity, based on coding region start and stop positions, recognizing that these do not necessarily correspond to the true transcriptional start and stop. An ANOVA model was constructed in JMP v.5.1 to explain variation in the length of 5′ upstream regions ($\log_{10}$ transformed) as a function of chromosome identity, gene density, and gene class (supplemental Table 1). Gene density for each gene was computed with a custom Perl script as the fraction of coding sequence in the 500-kb interval centered on that gene, arcsine-square-root transformed for analysis. The "spermatogenesis" gene expression class includes monocistronic genes categorized as spermatogenesis or "mixed spermatogenesis–somatic" from Reinke *et al.* (2004). Similarly, the germline category includes monocistronic "germline intrinsic" and "mixed oogenesis–somatic" genes from that study. The "operon" category used the first gene in an operon in calculations of upstream sequence length, regardless of its microarry-based expression classification from Reinke *et al.* (2004), and the "no data" category includes all remaining monocistronic coding loci. A separate ANOVA model was constructed with more specific expression categories, treating each classification from Reinke *et al.* (2004) separately (supplemental Table 2).

RESULTS

Operons are expressed in the germline of *C. elegans*: Previous microarray experiments identified 4106 genes with enriched expression in the hermaphrodite germline, relative to somatic tissues (Reinke *et al.* 2004). Of these, 1288 have peak expression during the fourth larval stage, specifically in germ cells undergoing spermatogenesis, while the remaining 2818 are expressed in germ cells at diverse stages of development, including undifferentiated and meiotic germ cells in larvae and adults and developing oocytes. These two groups of transcripts are referred to as "spermatogenesis enriched" and "germline enriched", respectively, in this report. *In situ* hybridization data confirmed that 98% of germline-enriched transcripts and 80% of spermatogenesis-enriched transcripts are indeed expressed in the germline (Reinke *et al.* 2004). We examined genes with germline-enriched or spermatogenesis-enriched expression for their incidence in operons. In total, 1075 of the 2818 genes with germline expression are located in operons (38%), even though only 15% of the genes in the genome reside in operons (2883 genes in 1150 operons) (Blumenthal *et al.* 2002). Conversely, operon genes on autosomes show germline gene expression about two- to threefold more frequently than expected at random ($P < 0.001$; Figure 1A). However, operons contain genes with spermatogenesis-enriched expression fivefold less frequently than expected ($P < 0.001$). Strikingly, both germline-enriched and operon genes are located very infrequently on the X chromosome (Reinke *et al.* 2000, 2004; Blumenthal *et al.* 2002), indicating that similar forces likely influence germline gene expression and operon organization on this chromosome.

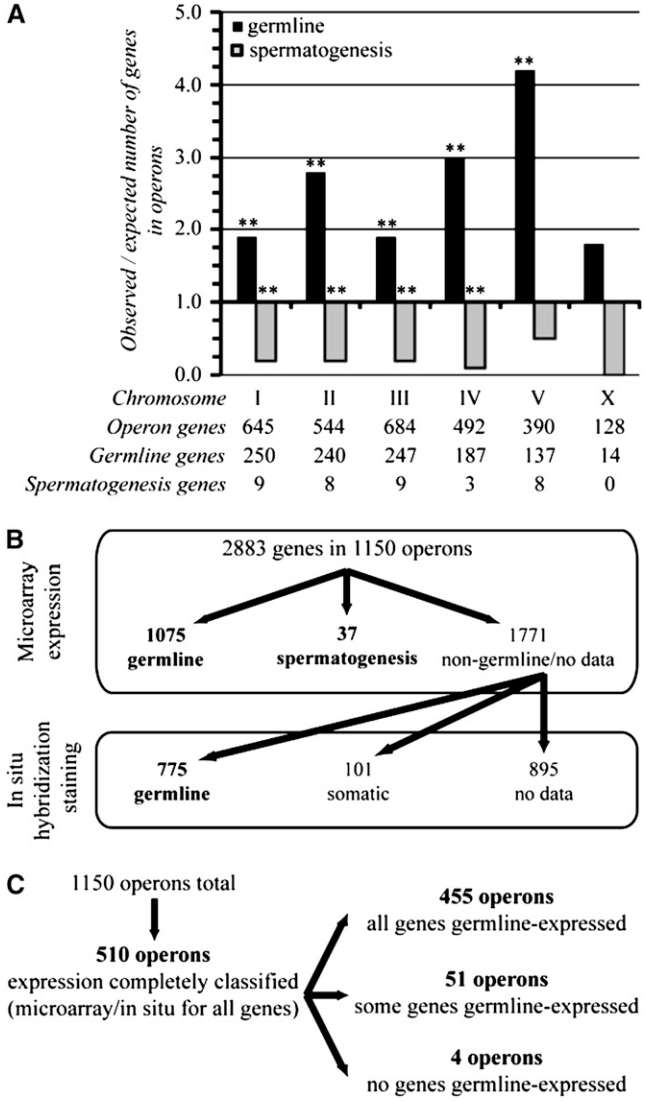

Figure 1.—Operons comprise germline-expressed genes. (A) Operon genes are expressed preferentially in the germline, but not during spermatogenesis. Per-chromosome hypergeometric probability test: $^{**}P < 0.001$ (either over- or underrepresented). Numbers listed at the bottom represent gene number in each category. (B) Nearly all genes within operons are expressed in the germline (1887 of 1988 genes with expression data). (C) Ninety percent of operons have all of their constituent genes expressed in the germline (455 of 510 operons with expression data for all gene members).

TABLE 1

Operons cluster along chromosomes

Chromosome	No. operons	No. clusters[a,b]	No. clusters[a,c]	No. operons in clusters[b]	No. operon genes in clusters[b]
I	192	36	12	129	332
II	166	26	22	86	226
III	203	38	4	142	370
IV	151	16	13	61	149
V	128	6	14	19	46
X	36	2	9	5	13

[a] 100-kb intervals with false discovery rate <0.05.
[b] Genomewide calibration.
[c] Per-chromosome calibration.

Overall, 37% of operon genes have germline-enriched microarray expression profiles, compared to 9% of monocistronic genes. However, this calculation likely underestimates germline expression of operons because the microarray experiments identify only transcripts enriched in the germline relative to somatic tissues (Reinke *et al.* 2004). Therefore, genes expressed in both the germline and soma, or genes not represented on the microarrays, would not be identified as germline enriched. Consequently, we used *in situ* hybridization patterns from NextDB (http://nematode.lab.nig.ac.jp/dbest/) to identify operon genes with germline expression for genes not previously defined as germline enriched in the microarray experiments (Figure 1B and supplemental data file 1). *In situ* data are available for 876 operon genes that lacked germline-enriched expression by microarray; of these, 775 (88%) have detectable germline staining. Nonoperon (monocistronic) genes that have germline-enriched microarray profiles also have germline staining *in situ* at similar frequency (98%), while only 20% of monocistronic genes without a germline-enriched microarray profile have germline staining (Reinke *et al.* 2004). The microarray and *in situ* hybridization data together indicate that as many as 95% of all genes in operons are expressed in the germline (Figure 1B, materials and methods).

We also determined the number of operons for which every gene within the operon showed expression in the germline (Figure 1C; supplemental data file 1). Five hundred ten operons have microarray data and/or visible *in situ* hybridization patterns for every gene within the operon. Of these, 455 operons are solely composed of germline-expressed genes; in some cases, these germline-expressed genes also show somatic expression. The remaining 55 operons contain at least one gene that shows somatic, but not germline, *in situ* hybridization staining. However, 51 of these 55 operons have at least one other gene that displays germline expression. For 99% of all 510 operons examined, at least one gene shows expression in the germline, while for 89% of these operons, every gene is germline expressed. However, detection of germline expression by *in situ* hybridization is easier than in somatic tissues, likely because of relative tissue size and RNA abundance. Expression in somatic tissues of most operon genes is therefore also likely, even if not detected with this method. Thus, expression in the germline appears to be an obligate, but not exclusive, characteristic of nearly all operons.

Operons cluster with monocistronic germline-enriched genes: In *C. elegans*, as in other organisms, neighboring genes in the genome often show similar expression in specific tissues; for instance, genes expressed in muscle or during spermatogenesis tend to cluster along chromosomes (Roy *et al.* 2002; Miller *et al.* 2004). We examined operon location and found that operons also tend to cluster along autosomes (Table 1; Figure 2A), and that they are rare on the X chromosome (Blumenthal *et al.* 2002). Specifically, when the genome was tiled into 100-kb windows, 74 windows showed a significant excess of operon genes within a given window than was expected from local gene density and chromosomal operon abundance. These 74 significant operon clusters compose 7.4% of the genome, yet harbor 22% of the operons in the *C. elegans* genome. A similar analysis that uses genomewide operon prevalence in the calculations, instead of per-chromosome operon abundance, identifies 124 significant clusters (12.3% of the genome) containing 38% of operons. Operons are more common in regions of high gene density (Spearman's $\rho = 0.27$, $P < 0.0001$), although the above analyses account for this effect, and significant operon clusters occur in regions of both high and low gene density. The median distance between operons is 36 kb, yet particular 100-kb intervals contain as many as 9 operons (*e.g.*, chromosome I, 4.16–4.26 Mb; Figure 2B). Control comparisons with somatic gene expression sets from intestine or neuronal tissues defined by tissue-specific gene expression profiling experiments (Fox *et al.* 2005; Pauli *et al.* 2006) revealed that monocistronic genes expressed in the intestine also cluster with operons, while neuronal genes do not.

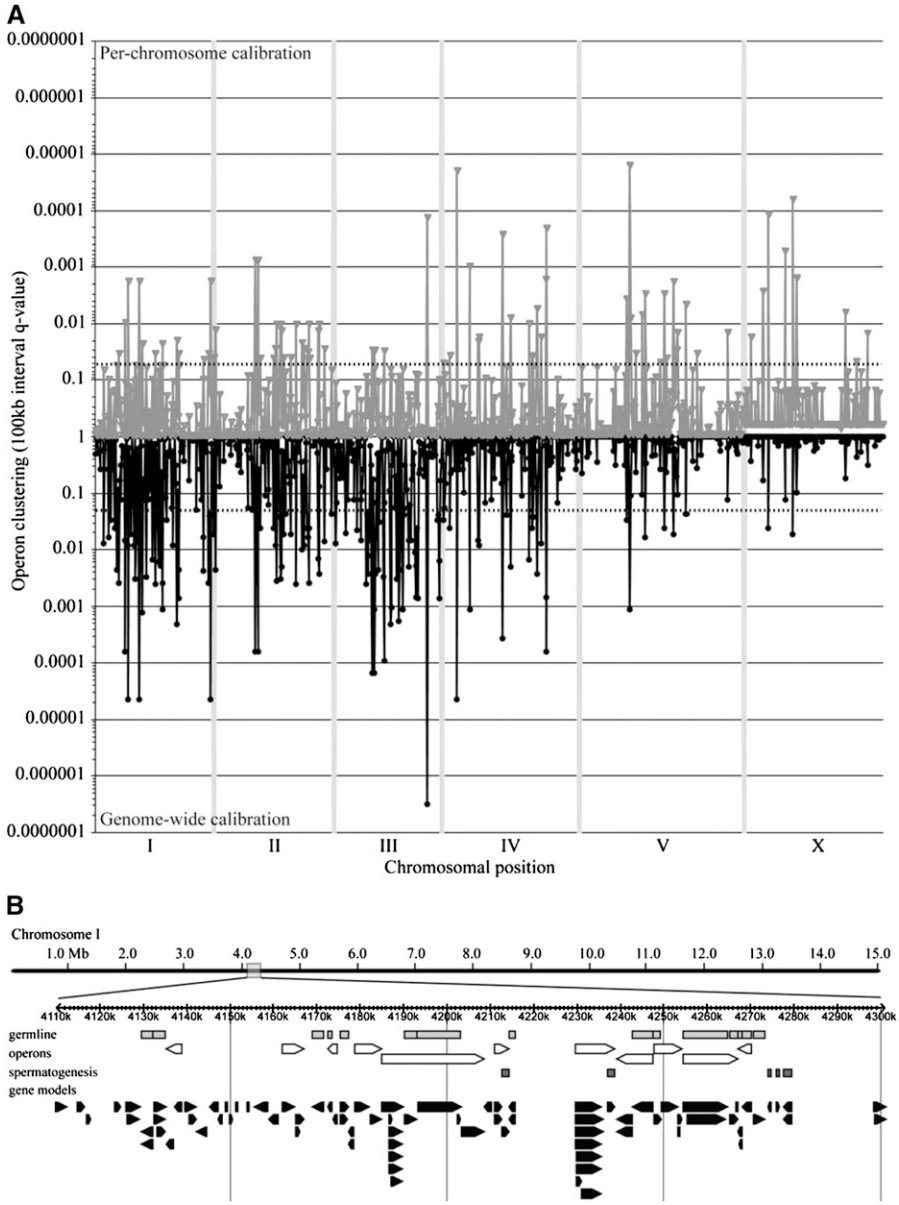

FIGURE 2.—Operons and monocistronic germline-enriched genes cluster along chromosomes. (A) False-discovery rate *q*-values for clustering of operons within nonoverlapping 100-kb intervals across the genome. Per-chromosome calibration (top half) assesses clustering given operon densities specific to each chromosome; genomewide calibration (bottom half) uses the genomic average proportion of genes that occur in operons. Horizontal dotted lines demarcate the $q = 0.05$ threshold, with smaller *q*-values indicating significant clustering in a given interval. Counts of significant clusters are summarized in Table 1. (B) An example of operons clustering with genes with germline-enriched expression along a ~200-kb region of chromosome I is shown, using the Genome Browser at WormBase (http://www.wormbase.org). Genes marked germline and spermatogenesis are defined from the microarray experiments only. This cluster of operons is somewhat unusual in that two spermatogenesis genes are present in operons. Additional examples from each chromosome are shown in supplemental Figure 2.

Notably, operon clusters will often include interspersed monocistronic genes with enriched expression in either spermatogenesis or in the germline, as determined by microarray analysis (REINKE *et al.* 2004) (Figure 3 and supplemental Figure 2). Monocistronic germline genes are significantly closer to operons than expected by chance (permutation test $P < 0.001$). The relative abundance of operons in 100-kb-long windows correlates positively with the proportion of monocistronic genes that have germline expression (Figure 3; Spearman's $\rho = 0.35$, $P < 0.0001$), further illustrating the physical co-occurrence of germline-expressed monocistronic genes with operons. The fact that operons cluster with monocistronic germline-expressed genes in small local regions along autosomes suggests that operons are coexpressed with monocistronic germline genes. Intriguingly, genes with spermatogenic expression are located in the same coexpression clusters as operons, yet generally are excluded from operons.

Germline expression influences operon composition more than housekeeping gene function: Genes within operons commonly encode proteins required for basic cellular "housekeeping" processes, including metabolism, transcription, and protein transport (BLUMENTHAL and GLEASON 2003), as do genes with germline-enriched expression (REINKE *et al.* 2004). The proportions of functional categories represented among genes in operons mirror closely the proportion of genes expressed in the germline and contrast with the representation of functional categories among genes with spermatogenic or somatic expression (supplemental Figure 1).

Possibly, genes are arranged in operons because the functions of the encoded gene products require or depend upon operon structure for proper regulation, as

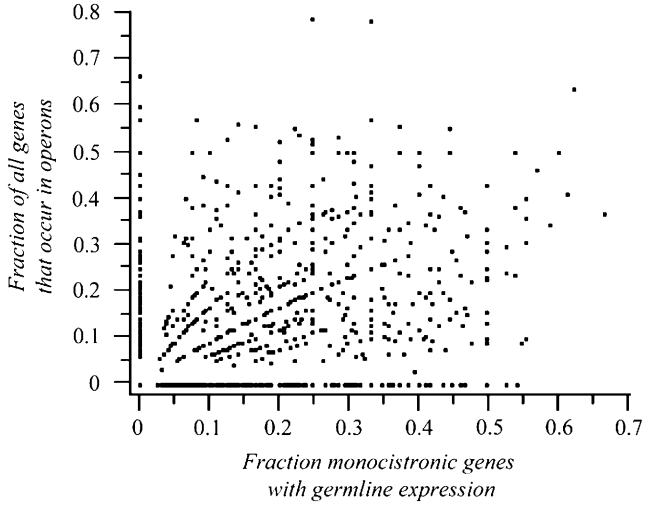

FIGURE 3.—Operon density correlates with the density of germline monocistronic genes. For nonoverlapping 100-kb intervals across the genome, the fraction of genes in each interval that occur in operons (operon density) is plotted against the fraction of genes that are both monocistronic and exhibit germline-enriched expression (density of germline-expressed monocistronic genes), on the basis of microarray expression data (Spearman's $\rho = 0.35$, $P < 0.0001$). The positive correlation indicates that genomic regions that have a high density of operons also tend to have a high density of germline-expressed monocistronic genes.

in bacterial operons. Alternatively, expression in the germline might facilitate operon origination or maintenance, independently of protein function. To distinguish between these two possibilities, we determined whether a stronger correlation exists between protein function and residence in an operon or between germline gene expression and residence in an operon. We independently defined sets of genes by selecting Gene Ontology category designations likely to reflect housekeeping functions, including "cellular metabolism," "cytoskeleton," and "protein transport." Taking cellular metabolism as an example, there are 685 genes annotated as such in the genome, 140 of which occur in operons (20%), representing a slight enrichment over the genome average of 15%. The frequency of metabolism genes is fairly uniform across all chromosomes, including the X chromosome (ranging from 2.8 to 4.3% of all genes per chromosome). However, metabolism genes in operons are rare on the X chromosome (3.6-fold underrepresented; Figure 4), even though metabolism genes overall are not rare on the X and not rare within operons. Similar patterns were found for genes in the cytoskeleton and protein transport categories, as well as in tissue-specific gene sets for the intestine and neurons (Figure 4; Fox *et al.* 2005; Pauli *et al.* 2006). Interestingly, intestine genes demonstrate a mild depletion from the X chromosome overall, which is in agreement with the fact that more of these genes are in operons. This example illustrates how restriction of genes from the X chromosome is peculiar to operons and germline-expressed genes, independently of whether they happen to encode proteins with metabolic or other housekeeping functions. Further, it suggests that housekeeping-gene function is not the driving force behind operon formation and/or maintenance. Notably, even though spermatogenesis and germline genes contain similar proportions of metabolism genes (see supplemental Figure 1), spermatogenesis genes are excluded from operons while germline genes are enriched in operons.

We also examined smaller sets of genes from a wide variety of functional groups defined by GO categories and found that germline gene expression is much more strongly associated with operons than is a specific protein function (Table 2). For example, genomewide, 26% of genes encoding protein phosphatases have germline expression, and 10% of all phosphatases lie within operons. If phosphatase function drives gene presence in operons, and not germline expression, then we expect only 26% of phosphatase genes in operons to be expressed in the germline. Instead, 90% of protein phosphatases encoded in operons exhibit germline expression (Table 2). Overall, examination of 16 different functional categories shows that operon genes have germline-enriched expression more frequently for members of all of these functional categories (Table 2; supplemental Figure 3 shows all categories; supplemental data file 2 has full data set). These data demonstrate that genes in operons are preferentially expressed in the germline independently of the molecular function of the encoded protein. Because expression in the germline is more strongly associated with operons than is the function of operon-encoded proteins, we conclude that germline expression is the key factor influencing operon formation and/or maintenance.

Operons have less 5′ regulatory sequence than monocistronic genes: Although expression in the germline is clearly a common property of almost all operon genes, many genes expressed in the germline are not found in operons. Why are certain genes preferentially located in operons? Possibly, operons accumulate genes with little temporal and spatial regulation of transcription and exclude genes with expression that is tightly regulated by sequence-specific transcription factors (usually during somatic differentiation). This hypothesis is supported by underrepresentation of genes from operons that encode tissue-specific differentiation factors and overrepresentation of genes encoding basic cellular machinery (Blumenthal and Gleason 2003). Furthermore, appropriate spatiotemporal expression for spermatogenesis genes is achieved by promoter-dependent processes, whereas germline genes rely primarily on their 3′-UTRs (Merritt *et al.* 2008), which is consistent with the observed exclusion of spermatogenesis genes from operons and the enrichment of germline genes.

We therefore wished to determine the relative complexity of the regulatory sequences of operon and non-

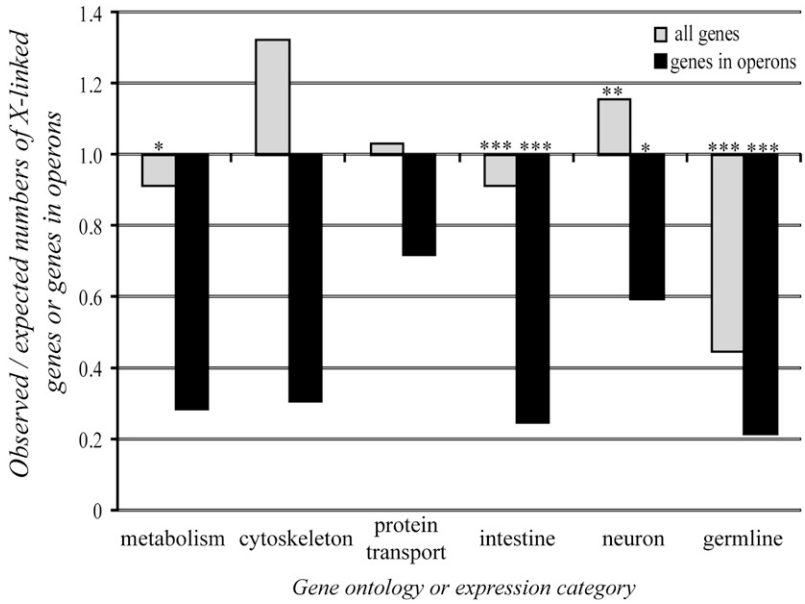

FIGURE 4.—Operons are biased for germline expression independent of protein function. The chromosomal distribution of gene sets defined by GO category (metabolism, cytoskeleton, transport) or by tissue-specific gene expression profiling (intestine, neuron, germline) was examined. Although all categories except the germline have as many genes on the X as expected, they are underrepresented for operons on the X, which is not expected if their protein function drives operon formation and/or maintenance.

operon genes. However, few *C. elegans* regulatory binding sites with cognate transcription factors have been identified. Additionally, transcription initiation sites are hard to define empirically in *C. elegans* because ~55% of mRNAs are *trans*-spliced to the SL1 splice leader at their 5′ ends, which removes some portion of the 5′-UTR (BLUMENTHAL and STEWARD 1997). We therefore used the length of noncoding regions upstream of the first exons as a proxy for 5′ regulatory complexity of a given gene. We found that the median length of regions upstream of operons is less than half that of monocistronic genes that lack germline expression (763 bp *vs.* 1536 bp; Table 3). Operon upstream regions also are significantly shorter than those of monocistronic genes that are associated with spermatogenesis (Table 3), despite the close physical proximity of these genes to operons. Both operon and monocistronic genes have shorter upstream regions when they occur in gene-dense portions of the genome relative to areas of low gene density ($P < 0.0001$) and the average length of upstream regions differs among chromosomes ($P < 0.0001$; supplemental Table 1), independently of gene expression profiles.

Intriguingly, the upstream regions of monocistronic spermatogenesis-expressed genes are significantly longer than those of monocistronic germline-expressed genes, suggesting that genes expressed during spermatogenesis might require more transcriptional regulation (supplemental Table 2). This hypothesis is supported by a recent gene-by-gene transgenic analysis of germline-expressed genes that demonstrates the reliance of spermatogenesis genes on their promoters for appropriate spatiotemporal expression, in contrast to germline genes that have dispensable promoters and are

TABLE 2

Operon genes are biased for germline expression, independently of function

	% germline expression of genes		% in operons of genes that are	
Functional class[a]	In operons	Not in operons	Germline expressed	Not germline expressed
Actin cytoskeleton	90	42	22	2
Cell adhesion	66	38	17	6
Chaperones	100	64	40	0
Golgi	100	81	28	0
GAPs	100	75	25	0
Isomerases	88	67	37	14
Mitochondrial	91	81	54	33
mRNA processing	95	90	40	22
Protein phosphatases	90	19	33	1
TAFs	100	75	50	0
Translation factors	100	90	41	0

[a] For all functional classes see supplemental data file 3.

TABLE 3

Upstream 5′ flanking regions of operons are shorter than most nonoperon genes

Expression category	*n*	Median[a]	Mean[a]	SD	Significance[b]
First genes in operons	1,106	763	1,866	2,990	A
Nonoperon germline	1,654	866	1,933	2,875	AB
Nonoperon spermatogenesis	1,185	985	2,174	3,230	B
Nongermline or no expression data	13,947	1,536	3,102	4,093	C

[a] Length in base pairs.
[b] Gene categories with different letters are significantly different with Tukey's honestly significant differences, accounting for the effects of gene density and chromosomal location (supplemental Table 1).

regulated by their 3′-UTRs (Merritt *et al.* 2008). Furthermore, post-transcriptional regulation of germline genes mediates appropriate spatial specificity of expression in artificial operon transgenes (Merritt *et al.* 2008).

Together, these observations are consistent with (1) spermatogenesis genes generally having greater capacity for upstream sequence to control their transcriptional regulation than germline genes and (2) a more limited role of transcriptional regulation to define spatial and temporal expression patterns for genes encoded in operons. Therefore, genes requiring less sequence-specific regulation—like germline genes, excluding spermatogenesis genes—might have greater freedom to enter operons by abandoning their individual promoters.

DISCUSSION

Operons constitute a major component of the organization of the *C. elegans* genome. Here, we present evidence that expression in the germline is an important contributor to the genomic distribution and composition of operons. We show that germline genes are found in operons at a significantly higher frequency than expected, and vice versa. Germline genes and operons also have the same chromosomal distribution, demonstrating a marked exclusion from the X chromosome. Monocistronic germline genes are clustered locally with operons, as well. Moreover, the upstream regulatory regions of both operons and germline genes are generally shorter than upstream regions of genes expressed in the soma. Finally, expression in the germline is more tightly associated with operon organization than is protein function or expression in other somatic tissues. Together these data indicate that expression in the germline shapes operon organization and genome distribution. However, the expression of operons is certainly not restricted to the germline. Most operons likely also have expression in somatic tissues, given that genes involved in basic cellular processes are commonly found in operons.

Evolution of operons: Both adaptive and parasitic models have been proposed for the evolution of the *trans*-splicing molecular machinery that is necessary for operon function (Lawrence 1999; Blumenthal and Gleason 2003). Regardless, given the ability of an organism to handle polycistronic transcripts, operon evolution has three components: (1) the origin of new operons from separate monocistronic genes, (2) the expansion of existing operons through the recruitment of monocistronic genes or the fusing of multiple operons, and (3) the maintenance of existing operon gene complexes in the face of mutational excision or deletion of genes from operons. The origin and expansion of operons may operate by similar processes, so we consider them jointly. The *trans*-splicing machinery in *C. elegans* provides a permissive environment for the formation of operons, but is not likely in itself to induce operon creation (Blumenthal and Gleason 2003). Selection for coregulation of multiple gene products from a single transcript also is an unlikely cause, as polycistronic mRNAs are rapidly *trans*-spliced following transcription (Spieth *et al.* 1993). Instead, we propose that limited regulatory requirements for the expression of genes in the germline, coupled with a weak reliance on promoters for cell-specific expression, facilitate their entry into operons.

A simple view of how an operon forms or grows in length, by joining two pieces of sequence in close proximity or by translocation, is through disruption of the 5′ promoter of the downstream portion, such that the most upstream promoter is then used for all members of the nascent operon, as well as through disruption of individual transcription termination signals. Qian and Zhang (2008) illustrate several examples in Caenorhabditis that are consistent with this idea. We propose that this process would facilitate the recruitment into operons of genes that have limited 5′ regulation necessary for their appropriate expression.

The near ubiquity of germline genes within operons, with their tendency for short upstream noncoding sequences and reliance on 3′-UTRs for proper expression (Merritt *et al.* 2008), is consistent with this model. The greater reliance on post-transcriptional regulation by germline genes simplifies the ability for such genes to share a promoter, because individual promoters do not provide much cell-type-specific information. The findings of Merritt *et al.* (2008) that 3′-UTRs of germline

genes induce cell-type specificity by blocking expression, rather than by inducing expression, also would facilitate the incorporation of germline genes into operons because any disruption of 3′-UTRs would result in a more ubiquitous pattern of expression that might be more selectively permissive than absence of expression. Moreover, the presence of internal promoters might impart a capability for unique somatic regulation of downstream genes in operons (Huang *et al.* 2007; Whittle *et al.* 2008). Loose clustering of monocistronic germline genes, perhaps due to chromatin architecture (Hurst *et al.* 2004), might also facilitate operon capture of germline genes that occur in close proximity, following deletion of short intervening sequences. The rarity of spermatogenesis genes in operons also is consistent with such a scenario, given their dependence on specific promoters for proper expression (Merritt *et al.* 2008) and generally longer upstream noncoding sequences that suggest greater 5′ regulatory complexity. This model does not inherently require natural selection to drive the creation or expansion of operons, although natural selection clearly is responsible for the evolutionary maintenance of particular operon structures (Stein *et al.* 2003; Qian and Zhang, 2008).

Exclusion of spermatogenesis genes from operons: Why are genes in operons expressed so frequently in the germline, but not during spermatogenesis? Operons and spermatogenesis genes frequently neighbor each other in the genome (Figure 2B), making it unlikely that global nuclear or chromatin architecture physically precludes spermatogenesis genes as candidates for incorporation into operons. It also seems unlikely that a property such as sustained, high levels of transcript synthesis would underlie inclusion in operons, because both oogenesis in hermaphrodites (and females) and spermatogenesis in males require this characteristic. Although *C. elegans* males are currently rare in nature (Barrière and Felix 2007), a relatively recent origin of hermaphroditism (Cutter *et al.* 2008) means that spermatogenesis gene expression in males is most relevant to their exclusion from operons over the course of evolution.

We hypothesize that sequence-specific transcriptional activation is required to drive expression of almost all genes expressed during spermatogenesis (and in somatic tissues as well), but many that are expressed in the germline require less sequence-specific transcriptional direction. We also hypothesize that a key factor permitting the inclusion of genes in operons is a low-complexity upstream regulatory region and weak promoter-dependent specification of expression, as seen for germline genes but not spermatogenesis genes (Table 3) (Merritt *et al.* 2008). A predisposition of germline genes to occur in operons also might feed back to inhibit the inclusion of spermatogenesis genes in operons. Because operons must have evolved in male–female ancestors of *C. elegans* and *C. briggsae* (Stein *et al.* 2003; Kiontke *et al.* 2004; Qian and Zhang 2008), expression of spermatogenesis and germline genes would need to occur in separate individuals rather than the common expression seen in the *C. elegans* hermaphroditic gonad. Thus, even a general germline operon promoter that might seem acceptable in a hermaphrodite would be inappropriate for spermatogenesis genes in ancestral species because of the requirement of male-limited expression.

We thank Yuji Kohara and members of his research team for generating and making public the large-scale *in situ* hybridization data. We also thank Tom Blumenthal for valuable discussions and advice and Jim Thomas, Amy MacQueen, and members of the Reinke lab for comments on the manuscript. This work was supported by a grant from the National Science Foundation (to V.R.). A.D.C. is supported by a National Science and Engineering Research Council Discovery Grant.

LITERATURE CITED

Barrière, A., and M.-A. Félix, 2007 Temporal dynamics and linkage disequilibrium in natural *Caenorhabditis elegans* populations. Genetics **176:** 999–1011.

Blumenthal, T., D. Evans, C. D. Link, A. Guffanti, D. Lawson *et al.*, 2002 A global analysis of *Caenorhabditis elegans* operons. Nature **417:** 851–854.

Blumenthal, T., and K. S. Gleason, 2003 *Caenorhabditis elegans* operons: form and function. Nat. Rev. Genet. **4:** 110–118.

Blumenthal, T., and K. Steward, 1997 RNA processing and gene structure, pp. 117–145 in *C. elegans II*, edited by D. L. Riddle, T. Blumenthal, B. J. Meyer and J. R. Priess. Cold Spring Harbor Laboratory Press, Plainview, NY.

Caron, H., B. Van Schaik, M. Van Der Mee, F. Baas, G. Riggins *et al.*, 2001 The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science **291:** 1289–1292.

Cohen, B. A., R. D. Mitra, J. D. Hughes and G. M. Church, 2000 A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat. Genet. **26:** 183–186.

Cutter, A. D., J. D. Wasmuth and N. L. Washington, 2008 Patterns of molecular evolution in Caenorhabditis preclude ancient origins of selfing. Genetics **178:** 2093–2104.

Fox, R. M., S. E. Von Stetina, S. J. Barlow, C. Shaffer, K. L. Olszewski *et al.*, 2005 A gene expression fingerprint of *C. elegans* embryonic motor neurons. BMC Genomics **6:** 42.

Huang, P., E. D. Pleasance, J. S. Maydan, R. Hunt-Newbury, N. J. O'Neil *et al.*, 2007 Identification and analysis of internal promoters in *Caenorhabditis elegans* operons. Genome Res. **17:** 1478–1485.

Hurst, L. D., C. Pal and M. J. Lercher, 2004 The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. **5:** 299–310.

Kiontke, K., N. P. Gavin, Y. Raynes, C. Roehrig, F. Piano *et al.*, 2004 Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proc. Natl. Acad. Sci. USA **101:** 9003–9008.

Lawrence, J., 1999 Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Curr. Opin. Genet. Dev. **9:** 642–648.

Lercher, M. J., T. Blumenthal and L. D. Hurst, 2003 Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. Genome Res. **13:** 238–243.

Merritt, C., D. Rasoloson, D. Ko and G. Seydoux, 2008 3′ UTRs are the primary regulators of gene expression in the *C. elegans* germline. Curr. Biol. **18:** 1476–1482.

Michalak, P., 2008 Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics **91:** 243–248.

Miller, M. A., A. D. Cutter, I. Yamamoto, S. Ward and D. Greenstein, 2004 Clustered organization of reproductive genes in the *C. elegans* genome. Current Biology **14:** 1284–1290.

Parisi, M., R. Nuttall, D. Naiman, G. Bouffard, J. Malley *et al.*, 2003 Paucity of genes on the Drosophila X chromosome showing male-biased expression. Science **299:** 697–700.

Pauli, F., Y. Liu, Y. A. Kim, P. J. Chen and S. K. Kim, 2006 Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in *C. elegans*. Development **133:** 287–295.

Qian, W., and J. Zhang, 2008 Evolutionary dynamics of nematode operons: easy come, slow go. Genome Res. **18:** 412–421.

Reinke, V., H. E. Smith, J. Nance, J. Wang, C. Van Doren *et al.*, 2000 A global profile of germline gene expression in *C. elegans*. Mol. Cell **6:** 605–616.

Reinke, V., I. San Gil, S. Ward and K. Kazmer, 2004 Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. Development **131:** 311–323.

Roy, P. J., J. M. Stuart, J. Lund and S. K. Kim, 2002 Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. Nature **418:** 975–979.

Spellman, P. T., and G. M. Rubin, 2002 Evidence for large domains of similarly expressed genes in the Drosophila genome. J. Biol. **1:** 5.

Spieth, J., G. Brooke, S. Kuersten, K. Lea and T. Blumenthal, 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. D., Z. Bao, D. Blasiar, T. Blumenthal, M. R. Brent *et al.*, 2003 The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. PLoS Biol. **1:** E45.

Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genome-wide studies. Proc. Natl. Acad. Sci. USA **100:** 9440–9445.

Wang, P. J., J. R. McCarrey, F. Yang and D. C. Page, 2001 An abundance of X-linked genes expressed in spermatogonia. Nat. Genet. **4:** 422–426.

Whittle, C. M., K. M. McClinic, S. Ercan, S. Zhang, R. D. Green *et al.*, 2008 The genomic distribution and function of histone variant HTZ-1 during *C. elegans* embryogenesis. PLoS Genet. **4:** e1000187.

Zorio, D. A. R., N. N. Cheng, T. Blumenthal and J. Spieth, 1994 Operons as a common form of chromosomal organization in *C. elegans*. Nature **372:** 270–272.

Communicating editor: B. J. Meyer