

Conserved elements associated with ribosomal genes and their trans-splice acceptor sites in *Caenorhabditis elegans*

Monica C. Sleumer¹, Allan K. Mah², David L. Baillie² and Steven J. M. Jones^{1,*}

¹Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, 570 W 7th Ave Suite 100, Vancouver, BC, Canada, V5Z 4S6 and ²Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6

Received July 28, 2009; Revised December 6, 2009; Accepted December 28, 2009

ABSTRACT

The recent publication of the *Caenorhabditis elegans* cisRED database has provided an extensive catalog of upstream elements that are conserved between nematode genomes. We have performed a secondary analysis to determine which subsequences of the cisRED motifs are found in multiple locations throughout the *C. elegans* genome. We used the word-counting motif discovery algorithm DME to form the motifs into groups based on sequence similarity. We then examined the genes associated with each motif group using DAVID and Ontologizer to determine which groups are associated with genes that also have significant functional associations in the Gene Ontology and other gene annotation sources. Of the 3265 motif groups formed, 612 (19%) had significant functional associations with respect to GO terms. Eight of the first 20 motif groups based on frequent dodecamers among the cisRED motif sequences were specifically associated with ribosomal protein genes; two of these were similar to mouse EBP-45, rat HNF3-family and *Drosophila* Zeste transcription factor binding sites. Additionally, seven motif groups were extensions of the canonical *C. elegans* trans-splice acceptor site. One motif group was tested for regulatory function in a series of green fluorescent protein expression experiments and was shown to be involved in pharyngeal expression.

INTRODUCTION

Regulatory sequences are conserved both across different species and across different genes within a single species.

Recently, a study was published in which more than 100 000 short conserved regions in *Caenorhabditis elegans* upstream regions were identified by comparison to other nematode genomes (1). We have now pursued those findings by determining which of the conserved sequences are found repeatedly throughout the *C. elegans* genome.

The cisRED database contains 158 017 conserved motifs in the upstream regions of 3847 *C. elegans* transcripts, representing 3458 different genes (1). These motifs were identified by comparing the *C. elegans* upstream regions to their orthologous counterparts in seven other nematode genomes (*C. briggsae*, *C. remanei*, *C. brenneri*, *C. japonica*, *Pristionchus pacificus*, *Brugia malayi* and *Trichinella spiralis*). Upstream regions were defined as the intergenic region or 1500 bp, whichever was shorter. Conserved motifs in each orthologous upstream sequence region set were identified using the motif discovery algorithm MotifSampler (2).

In order to identify which motifs were similar to previously characterized transcription factor binding sites (TFBSs) from other species, all motifs were scanned using binding models from the ORegAnno (3), JASPAR (4) and TRANSFAC (5) databases. Each motif was assigned a *P*-value indicating the significance of its similarity to any applicable TFBSs; 26% of the motifs were found to be significantly similar to at least one known TFBS. However, this result was limited by the set of transcription factors (TFs) for which illustrative binding sites existed—a relatively small set compared to TFs with unknown binding sites. Additionally, for those few TFs for which we had binding sequence data, the mechanism of TF binding is not understood well enough to accurately distinguish true binding sites from similar sequences. The significance and function of most motifs remains unknown, and we anticipate that many of the remaining motifs may represent previously undiscovered regulatory elements.

*To whom correspondence should be addressed. Email: sjones@bcgsc.ca

Each TF typically binds to multiple sites in the genome and is involved in the regulation of expression of several genes. The goal of this research was to find groups of genes that have similar sequences in their upstream regions. We hypothesized that such a group of genes might be related in terms of function, gene regulation, and gene expression. This hypothesis was tested by examining functional information in the Gene Ontology (GO) (6) and biochemical pathway databases, and by testing the influence of the sequence on the expression of the genes *in vivo* using green fluorescent protein (GFP) expression constructs.

It is important to note that we initially assumed that shared sequences in upstream regions were regulatory elements that functioned as TFBSs at the DNA level. However, sequences identified in this manner could also function at the RNA level as unannotated ncRNA genes (7), RNA-binding protein binding sites (8), RNA secondary structure motifs (9) and microRNA targets (10,11). The mechanism of the function of the identified sequences is a separate question that was beyond the scope of the research described here.

Given the large quantity of available genomic sequence and gene expression data, studies in which these data are considered simultaneously are numerous. Such studies typically focus on a small set of genes that are coexpressed in the same tissue type or upregulated under specific environmental conditions. For example, the algorithms PhyloCon (12) and PhyloGibbs (13) can take the upstream regions of a set of coexpressed genes plus their orthologs in other species and produce a set of motifs that are both overrepresented in the coexpressed gene set and conserved in the orthologous sequences. PhyloCon was subsequently used by Zhao *et al.* (14) to discover several regulatory motifs in *C. elegans* and *C. briggsae* that were responsible for muscle expression.

Other computational methods to discover regulatory elements have focused on modules: sets of motifs that tend to co-occur. Regulatory module detection algorithms typically take a set of coexpressed genes and a list of TF-binding models and search for genes in the set that have two or more of the motifs in their upstream regions in close proximity to each other. Many of the programs also consider orthologous conservation; see Van Loo and Marynen (15) for a detailed review. We did not find other analyses in which orthologous sequence was initially used to find conserved genomic regions, which were then compared to each other to find common subsequences in a genome-wide manner.

We used DME to compare cisRED motif sequences to each other and determine which subsequences were conserved more often than expected among nematode upstream regions. DME is an advanced deterministic word-counting motif discovery algorithm that generates an exhaustive count of all *n*-mers in both a foreground sequence set and a background sequence set (16). It then determines all sequences and sequence variations that occur more frequently than expected in the foreground set given their frequency in the background set and the relative sizes of the two sets. It has previously been used in a number of applications, including tissue-specific

promoter analysis (17,18), analysis of conserved noncoding elements in mammalian genomes (19), motif discovery in splice junctions of tissue-specific RNAs in humans (20), DNase hypersensitive site analysis (21), promoter analysis of specifically expressed genes identified by chromatin immunoprecipitation (ChIP-chip) (22) and TF-specific promoter analyses (23).

The above-mentioned data sets analyzed by DME feature numerous segments of promoter sequences of which some may contain a motif in common with other sequences, some contain multiple instances of the same motif or more than one motif, while other sequences contain no TF-binding motifs. These similarities between the cisRED motif sequences and data successfully analyzed by DME are what lead us to apply DME to the current problem of analyzing common sequences among cisRED motifs.

In this study, we have attempted to find and characterize novel functional elements by identifying sequences found repeatedly among the identified cisRED motifs. Using all 158017 motifs from the cisRED *C. elegans* database of conserved elements, we employed the DME algorithm to place the motifs into groups in which all members of a group contained the same or similar subsequence. In this case, the foreground set consisted of the *C. elegans* sequences of all cisRED motifs, and the background set consisted of all *C. elegans* upstream regions from which the motifs derived. DME is a deterministic algorithm that will always produce the same result; therefore, we used it in an iterative manner in which the central bases of all members of each motif group were masked out before the next motif group was found. DME requires user input for the width of the overrepresented sequence to be found and the degree of conservation; we ran DME independently at the five most common widths of the cisRED motifs (6, 8, 10, 12 and 14 bp) and decreased the conservation requirement as the width increased. This strategy formed groups of motifs based on sequence similarity; many of the motifs belonged to more than one group.

We then used the web-based tool DAVID (24) and the command-line tool Ontologizer (25) to assess the significance of each motif group by determining whether genes that shared members of the same motif group also shared GO and other annotations. We found that 19% of the motif groups were associated with annotations such as nucleotide binding, transferases and transit peptides. The most common association among the motif groups was with ribosomal protein genes; we observed that eight of the first 20 motif groups of 12-bp width were significantly associated with ribosomal genes, so we focused further research on these.

The most significant ribosome-associated motif group was associated with 120 genes, of which 28 were ribosomal and others were involved in embryonic development, larval development and multicellular organismal development. For 11 of the 120 genes, we obtained GFP construct expression data from the BC *C. elegans* Gene Expression Consortium (26). We tested the importance of the motif for the expression pattern of each of the 11 genes using a series of GFP constructs. Four of the 11 genes showed a

difference in expression between constructs including the motif and constructs excluding the motif or with a mutated motif. The motif appears to be necessary for pharyngeal expression of these genes.

METHODS

Motif grouping with DME

The *C. elegans* sequence of all cisRED motifs were extracted and combined into a single FastA file. All *C. elegans* cisRED upstream regions were combined into a background FastA file (original genome build: WS170). A version of DME that did not preface the word-counting step with a repeatmasking step and did not weigh motif information content (IC) by base composition was obtained from Dr Andrew Smith. DME was run using the parameters indicated in Table 1 at each width. After each iteration, the two central bases of each motif in the new group were masked to Ns and DME was re-run until it either could no longer find any overrepresented sequences, or reached 900 iterations. All motif groups were uploaded to the cisRED database as *de novo* motif groups.

Functional characterization with DAVID

Entrez gene IDs for all genes associated with the first 20 groups at each width were extracted and analyzed by DAVID via the HTML-based API. The following annotation categories were included in the HTML

Table 1. DME Parameters

Width	Parameters
6	-C 0.25,0.25,0.25,0.25 -n 1 -i 2.0 -w 6 -r 0.25 -g 0.0
8	-C 0.25,0.25,0.25,0.25 -n 1 -i 1.8 -w 8 -r 0.25 -g 0.0
10	-C 0.25,0.25,0.25,0.25 -n 1 -i 1.7 -w 10 -r 0.25 -g 0.0
12	-C 0.25,0.25,0.25,0.25 -n 1 -i 1.6 -w 12 -r 0.25 -g 0.5
14	-C 0.25,0.25,0.25,0.25 -n 1 -i 1.5 -w 14 -r 0.25 -g 1.0

Parameters used for DME at each of the five widths are indicated.

Table 2. Primers used for generation of GFP constructs

Gene	Primer including motif	Primer mutating motif	Primer excluding motif	Reverse primer
<i>C13B9.3</i>	cggggagggtctcgcaacgaaatga	cggggagggtctcttaacgaaatga	ttcactggttctcgttggga	cggcgatcaacacgattg
<i>C26D10.2</i>	ttacttcgctcgagaccatacga	ttacttcgctaagagaccatacga	cgaatgggtatcgtttcgc	gtgttcctctcgattcgaaa
<i>C34E10.6</i>	tccatttcgttgcgagaccgctg	tccatttcgtaagagaccgctg	gcggtctagcctgtttcagt	taacgaacgcgaagcgata
<i>F07A11.2a</i>	tctcaaccggagcgttgcgagacc	tctcaaccggagcgttaagagacc	tgatcttcgatcgttctcg	aattccgcagatttggatg
<i>F09B9.3</i>	agacgaacatcgtcgcgagaccag	agacgaacatcgtcgaagaccag	ggacgaatagctcgcatctc	tctgcgttatggaagaacagaa
<i>F25H2.5</i>	aggtcgggtctcgcacagtgctgaaatga	aggtcgggtctctcagctgctgaaatga	tcgtttcatttgcgtcggag	tcagttgctgattttccg
<i>F54D8.2</i>	atttcaccggctggtctcgcagcgaa	atttcaccggctggtctcttagcgaa	agaccgctctcctgttattt	cgggtgatgctggataacct
<i>M01F1.3</i>	attgcgtatcgtgagacccat	attgcgtatcgtgaaagacccat	atggcttttcgctatcct	accgagctaggatgcttaaa
<i>T05H4.1</i>	acttctgagcgttgcgagaccgtg	acttctgagcgttaagagaccgtg	tcacaaaagaacacctccc	ttgatatcgtcattctgttggag
<i>Y48G8AL.8a</i>	acacaagatcgcgagaccat	acacaagatcgcgaagagaccat	tcgcttgcgcctttaaata	gtgaacctctgatttcgac
<i>Y57G11C.13</i>	tcgatcgcgcaaacccgtctcgaaa	tcgatcgcgcaaacccgtctcgaaa	aaaccgctcctgaaactg	tctgaaatattgatgttgaatgag

Primers for the three GFP constructs generated for each of the 11 genes. All constructs used the same reverse primer near or overlapping the ATG of the tested gene; this was also the same reverse primer that was used by the BC *C. elegans* Gene Expression Consortium. The 'Primer Mutating Motif' differs from the 'Primer Including Motif' by two bases; the same mutation was introduced in all cases; the 'Primer Excluding Motif' was 12–62 bases downstream of the 'Primer Including Motif'.

links: GOTERM_BP_ALL, GOTERM_CC_ALL, GOTERM_MF_ALL, INTERPRO.PFAM, PIR_SUPERFAMILY, KEGG_PATHWAY, SP_PIR_KEYWORDS, BIND, DIP and MINT. For motif groups with more than 400 associated genes, only the first 400 genes were analyzed (in numerical order) due to limitations imposed by the DAVID interface.

Functional characterization with Ontologizer

Wormbase IDs for all genes used in the cisRED analysis pipeline were extracted as a background list. Wormbase IDs associated with each motif group were also extracted. GO annotation files (gene_association.wb and gene_ontology_edit.obo) were downloaded from www.geneontology.org in October 2008. Ontologizer was run using the command-line Java jar using the following parameters: -g obo_file -a association_file -p background_list -s foreground_list -m 'Benjamini-Hochberg' -c 'Term-For-Term' -i.

GFP constructs

Primers were designed for GFP constructs as shown in Table 2. Constructs were generated by polymerase chain reaction (PCR) and injected into the gonads of gravid hermaphrodites. DNA constructs were generated via fusion PCR as previously described by Hobert (27) using DNA template prepared from N2 genomic DNA (Bristol, Baillie Laboratory strain BC49). Phusion polymerase (Finnzymes, New England Biolabs Cat: F530) was used for all PCR reactions to ensure fidelity of the resultant construct. Promoter-containing sequences were fused upstream of the GFP coding region from GFP-coding cassette *pPD95.67*. The reverse-promoter-associated primer includes a segment complementary to the forward primer used for amplification of the GFP-reporter cassettes. The primers used for amplification of the GFP-encoding region are as follows: GFP D-AAG GGC CCG TAC GGC CGA CTA GTA GG, GFP D*- GGA AAC AGT TAT GTT TGG TAT ATT GGG and GFP C- AGC TTG CAT GCC TGC AGG TCG ACT (26–28).

All nematode microinjections were conducted using either Olympus BH2-HLSH or Zeiss 473016 inverted microscopes. PCR constructs were injected at an average final concentration of 30 ng/μl, along with 100 ng/μl of the marker construct, *pCeh361 [dpy-5(+)]* (29), into the syncytial portion of the somatic gonad of target strain *dpy-5(e907)* (CB907). *Dpy-5* rescued wild-type F1s were individually plated. Wild-type F2 lines were selected to establish the transgenic lines. Wild-type F2 were analyzed.

Microscopy

A Zeiss AxioScope equipped with a QImaging camera and the appropriate optical filter sets was used for GFP expression pattern analysis. Worms were immobilized on moist agarose pads (2% in water) with 5 μl of 100 mM sodium azide (in water) immediately prior to imaging. All images were taken with identical filter, lens and camera settings for all image sets. Images were exposed for three seconds and captured using QCapture software.

Results

Motif grouping

Motifs were formed into 3265 groups based on sequence similarity. DME found sequences that appeared more often than expected in the set of cisRED motif sequences compared to the 3847 upstream regions from which the motifs were derived. We used DME in an iterative process; the most significant motif group was found first, and then the central two bases of each instance of the group were masked with Ns before the next iteration. In this manner, we ensured that each group was unique and not just a minor variation or 1-bp shift of a previously found motif.

DME iterations were run independently for widths of 6, 8, 10, 12 and 14 bp, the same primary widths as the cisRED motifs themselves. All 158 017 of the motifs in the cisRED database had a width of at least 6 bp, therefore, they were all eligible for grouping at that width; in contrast, only 72 935 of the cisRED motifs had a width of at least 14 bp (Table 3).

The parameters for DME were set such that the stringency for motif sequence similarity was relaxed as the motif width increased. The requirements for width-six motif groups were set to a maximum stringency: an IC of two or perfect match. This meant that all members of motif groups of width six contained the same hexamer. For motif groups of width eight, the IC requirement was 1.8, meaning that each motif group consisted of motifs containing octamers that differed in only one position. For motif groups of width 14, the IC requirement was only 1.5. In spite of the relaxed IC requirements, motif groups of high width were far smaller than those of low width. This was due to the relative rarity of long similar sequences, and the smaller set of motifs from which to draw.

Seventy-six overrepresented hexamers were found among the cisRED motifs. The first motif group, named 6-0, was the hexamer GATAAG. This sequence appeared 1469 times among the *C. elegans* sequences of the cisRED motifs and only 2245 times among all *C. elegans* upstream regions from which the cisRED motifs were derived. As the cisRED motif sequences amounted to a total of 2.15 Mbp and the upstream regions amounted to a total of 4.16 Mbp (non-repeatmasked), the expected number of occurrences for the hexamer among the motifs was 1160; therefore, this hexamer was significantly overrepresented (P -value 1.9E-39 by the binomial test). Similarly, the second hexamer motif group (named 6-1) consisted of motifs containing the palindrome CACGTG. This sequence appeared 1398 times among the *C. elegans* upstream regions, of which 1006 were conserved (P -value 8.4E-54). After 76 hexamer-based motif groups were found, DME was no longer able to find any further hexamers that were overrepresented among the cisRED motifs with respect to the upstream regions. Of the 76 hexameric motif groups, 45 (59%) had P -values <0.05. Group 6-75, the last hexameric motif group, was the sequence GCGTTG, which appeared 813 times in total among the upstream regions and 412 times among the cisRED motifs, only slightly more often than the expected number of 407 (P -value 0.75). The smallest hexamer motif group (6-62) was the

Table 3. Summary of motif grouping results

Width (bp)	Num of available motifs	Min IC	Num of groups	Smallest	Largest	Num of motifs in groups	Group ID range
6	158 017	2.0	76	130	1714	28 478	1–76
8	146 052	1.8	489	4	545	19 824	77–565
10	125 822	1.7	900	7	282	16 583	566–1465
12	101 348	1.6	900	7	155	11 963	1466–2365
14	72 935	1.5	900	5	91	8725	2366–3265

DME was run iteratively on cisRED motifs to form them into groups based on sequence similarity; DME was run independently at widths 6, 8, 10, 12 and 14 bp. 'Num of Available Motifs' shows the number of cisRED motifs that met the width requirement. 'Min IC' shows how the required information content decreased as the width increased. 'Num Groups' shows how many iterations were run, and therefore how many motif groups were generated, at each width. For widths 6 and 8, DME terminated automatically with no motif groups left to find after 76 and 489 iterations respectively, while for widths 10, 12 and 14, the process was terminated after 900 iterations. 'Smallest' and 'Largest' show the number of motifs in the smallest and largest group of each width, respectively. 'Num Motifs in Groups' shows the number of eligible motifs in at least one group after all iterations of that width. Integer identification numbers ('Group IDs') were assigned sequentially to each motif group to identify it in the cisRED database; the range of group ID numbers for groups at each width is shown.

palindrome GGGCCC, which appeared 130 times among cisRED motifs and 236 times in the upstream regions (P -value 0.17). After 76 iterations, 28 478 (18%) of motifs were in at least one hexamer motif group (Table 3).

Four hundred eighty-nine overrepresented octamers were found. Of the cisRED motifs, 11 965 were of width 6 or 7 and, hence, were ineligible for width-eight motif grouping, leaving 146 052 available motifs (1). The first motif group of width 8 (named 8-0) consisted of motifs that contained the sequence CTGCGYCT. This sequence appeared 506 times in total among the upstream regions and 371 times among cisRED motifs (the expected number of occurrences among the cisRED motifs was 261; P -value $2.5E-23$). Members of the second octamer motif group (8-1) all contained the sequence GHGCGCGC. This sequence is a partial palindrome (palindromic whenever the base in the second position is a C). It is also possible for instances of this sequence to overlap substantially with each other. Including all overlapping instances on both strands, DME counted a total of 447 instances of this sequence among the upstream regions, of which 328 were in cisRED motifs (P -value $6.5E-21$). The smallest octamer motif group (8-390) was the palindrome GGCTAGCC, which appeared four times among cisRED motifs and only six times in total among the upstream regions (P -value 0.38).

After 489 motif groups were found and their central bases were masked out, DME was unable to find any more octamers that were overrepresented among cisRED motifs. Of the octameric motif groups, 122 had P -values <0.05 . Group 8-488 (the last octamer motif group) was the sequence TGACGGTG, which appeared 73 times in the upstream regions, of which 37 were in cisRED motifs; the expected number of occurrences among cisRED motifs was 36 (P -value 0.62). Many of the octamer motif groups overlapped with each other and with the hexamer motif groups; 19 824 of the motifs were in at least one octamer motif group.

Nine hundred overrepresented decameric motif groups were found. Of the entire set of cisRED motifs, 125 822 were at least 10 bp wide and therefore eligible for decameric grouping. The first width-10 motif group (named 10-0) had the consensus sequence GAKACGCAGN; because the last position is held by all four bases, this pattern is a nonamer rather than a decamer. Sequences that matched this pattern appeared 402 times in the upstream regions, of which 282 were within cisRED motifs (P -value $4.5E-14$). Group 10-1, the second decameric motif group, had the consensus sequence GTCYCGCMRC, which appeared in 155 cisRED motifs and 215 times in the upstream sequences (P -value $2.1E-11$). There were five smallest decamer motif groups, all of which appeared seven times in both the upstream regions and the cisRED motifs (iterations 789, 790, 811, 812 and 813; P -value 0.0099). DME did not terminate automatically and was stopped after 900 iterations had run, and 819 of the groups had a P -value <0.05 . The 900th decameric motif group, 10-899, had the consensus sequence TAACMCGWCT; this sequence appeared 14 times in the upstream regions, 11 of which

were in cisRED motifs (P -value 0.038). After 900 iterations, 16 583 of the motifs were in decameric motif groups.

Nine hundred overrepresented dodecameric motif groups were found. There were 101 348 cisRED motifs with a width of 12 bp or greater. Eight of the first 20 width-12 motif groups were very interesting and are discussed in detail below. The three smallest dodecameric motif groups were 12-658, 12-663 and 12-670, all of which appeared seven times among both the cisRED motifs and the upstream regions (P -value 0.0099). Just as for the decameric motif groups, DME did not terminate automatically and was stopped after 900 iterations; motif group 12-899 had the consensus sequence YGGCGGC RSCAB. Sequences matching this pattern were observed 12 times in the upstream regions and 11 times among cisRED motifs (P -value 0.0045). After 900 iterations, 11 963 of the motifs were in dodecameric motif groups; all but one had a P -value of <0.05 .

Nine hundred overrepresented width-14 motif groups were found. Only 72 935 (46%) of the motifs were wide enough to be eligible for width-14 motif grouping. Motif group 14-0, the first width-14 group, overlapped very strongly with group 12-0, the first width-12 group. However, it was not a superset, because group 12-0 included some motifs that were exactly 12 bp wide, and it was also not a subset because the IC of group 14-0 was lower and more sequence variations were included. Group 14-0 had a consensus sequence of KSGTCYSSMRCGA, which appeared 136 times in the upstream regions and 91 times among cisRED motifs (P -value 0.00024). The second group (14-1) had the consensus sequence RYRWGTGYKASYGT, which appeared 44 times in the upstream regions and 37 times among the cisRED motifs (P -value $7.4E-6$). The four smallest motif groups, 14-545, 14-877, 14-879 and 14-884, all appeared five times among the cisRED motifs. The first three of these also appeared five times among the upstream regions (P -value 0.037), while 14-884 appeared six times among the upstream regions (P -value 0.13). After 900 iterations, DME was terminated. Group 14-899 had the consensus sequence RWMAWTMTYGKCGT, which appeared six times among the upstream regions, all of which were in cisRED motifs (P -value 0.019). Of the eligible motifs, 8725 were in groups after 900 iterations. All but six width-14 motifs had P -values of <0.05 .

In total, 45 312 (29%) of motifs were in 3265 overlapping groups after all DME iterations were completed. Fifty percent of the motifs in a group were in more than one group. Nine of the groups (two octameric, four decameric and three dodecameric) had an N at the beginning or end of the consensus sequence, indicating that the true pattern was one base shorter. A summary of the motif grouping result numbers is shown in Table 3. Motif groups are browsable via the cisRED web interface at: www.cisred.org/c.elegans4/all_groups?showab=0&showdn=1. Additionally, each motif group has its own URL at: www.cisred.org/c.elegans4/group_content_view?aid={Group_ID} (note

that '{Group_ID}' must be replaced by the Group ID of the motif group in question).

Functional characterization of genes associated with motif groups

We used DAVID (24) to perform a preliminary examination of the associated genes of each of the first 20 motif groups at each width to see if they had any functional similarities with respect to annotations in the GO, the Protein Information Resource (PIR) (30), and other gene annotation sources. *P*-values were adjusted for multiple testing correction using the Benjamini–Hochberg method in the DAVID web interface. For the hexameric motif groups, all 20 groups had significant multiple testing-corrected associations. For example, the fifth group, consisting of all instances of the sequence GGGCGG among the motifs, was significantly associated with genes involved in DNA binding, including homeobox genes. Six of the first 20 motif group associations were with ribosomal proteins.

For the octameric motif groups, 17 out of the first 20 groups had significant associations. For example, group 8-0 was associated with ATP-binding and mitochondrial proteins. Four of the top 20 motif groups were associated with ribosomal proteins. Similarly, 16 of the first 20 dodecameric motif groups were significantly associated with gene categories such as nucleotide binding, cytoplasmic proteins, transit peptides and anatomical structure development. Once again, eight of the dodecameric groups were associated with ribosomal proteins. The same general results were observed for dodecameric and width-14 motif groups: 11 of the dodecameric groups were significant (eight of them with ribosomal proteins), and eight of the width-14 groups were significant, six of them with ribosomal proteins.

In order to obtain a more complete picture of the significance of all motif groups with respect to associated gene function, we used Ontologizer to analyze all groups for overrepresented GO terms with respect to the background set of 3458 genes. Out of the background set, 2789 of the genes were annotated with at least one GO term. We used a Benjamini–Hochberg multiple testing correction for each gene list that was associated with a motif group. It is also possible to apply a further Bonferroni-corrected *P*-value threshold of 1.5E-5 (0.05/3265) to account for the 3265 gene groups that we analyzed. Of the 3265 motif groups tested, 612 (19%) had a Benjamini–Hochberg corrected *P*-value of 0.05 or less for at least one overrepresented GO term, and 26 of these had a Benjamini–Hochberg-corrected *P*-value of 1.5E-5 or less.

Description of eight motif groups associated with ribosomal genes

We observed that many of the significant motif groups were associated with ribosomal genes, so we decided to concentrate further research on these motifs. Specifically, eight of the top 20 dodecameric motif groups were associated with between six and 28 ribosomal genes, with a total of 63 ribosomal genes between them

(Table 4). All eight of them were found to have significant associations with ribosomal genes by DAVID (seven by Ontologizer as explained in the 'Discussion' section). There were only 96 ribosomal genes in the set used for the cisRED database, and there are only 176 ribosomal protein genes in total in the *C. elegans* genome, so this was a substantial proportion. Three pairs of ribosomal proteins were on bidirectional promoters and therefore had the same upstream region. Each ribosomal gene had no more than one instance of each motif group (with one exception: there were two instances of motif group 12-3 upstream of *Y119D3B.16*) and no more than three different motif groups in its upstream region.

The first dodecameric motif group, 12-0, had the most significant *P*-value with respect to ribosomal genes of the eight motif groups and also had the most members. We observed that it was GC-rich (75% GC) and very strongly conserved—the IC of all 120 instances of the group was 1.7 and the IC of the 28 instances near ribosomal genes was 1.8. It tended to appear ~300 bp upstream of the translation start site (ATG) of the gene and was not strand-biased (Figure 1). It also tended to occur about 30 bp upstream or downstream of one of the other ribosomal motifs such as 12-5, 12-11 or 12-18. It was not found to co-occur in an upstream region with motif group 12-3 or 12-4. Group 12-0 was found upstream of 28 ribosomal genes, of which two pairs of genes were on bidirectional promoters and therefore had the same upstream regions: *lsm-1* (*F40F8.9*; a small nuclear ribonucleoprotein splicing factor) and *rps-9* (*F40F8.10*), and *rps-30* (*C26F1.4*) and *rpl-39* (*C26F1.9*).

Four of the ribosomal motif groups were similar to known TFBSs in other species. None of the ribosomal motif groups overlapped significantly with motif annotations from the cisRED database. We used the Transcriptional Element Search System (TESS) (31) to determine whether any of the ribosomal motif groups were similar to TFBSs beyond those that were already annotated. Motif group 12-1 was found to be significantly similar to the binding site for mouse ZF5, whose consensus binding sequence is GSGCGCGR (32) (Table 4). Motif group 12-3 had significant similarities to the binding sites for both EBP-45 [binding sequence TGTTTGC (33)] and HNF3-family TFs [binding consensus sequence described as YGTTTTRT in rat (34) and TRTTTGY in the frog *Xenopus laevis* (35)]. Motif group 12-8 was found to be significantly similar to the binding site for Delta EF1 in the chicken genome [binding sequence AGGTG (36)] even though the motif group sequence was not a perfect match. Motif group 12-18 was significantly similar to the Zeste binding site in *Drosophila melanogaster* [binding consensus sequence YGAGYG (37)].

The motif groups are associated with cytoplasmic ribosomal proteins. *Caenorhabditis elegans*, like other eukaryotes, has ribosomes associated with two different intracellular localization patterns, cytoplasmic and mitochondrial (38). We examined whether the ribosomal motif groups were associated with cytoplasmic ribosomal genes,

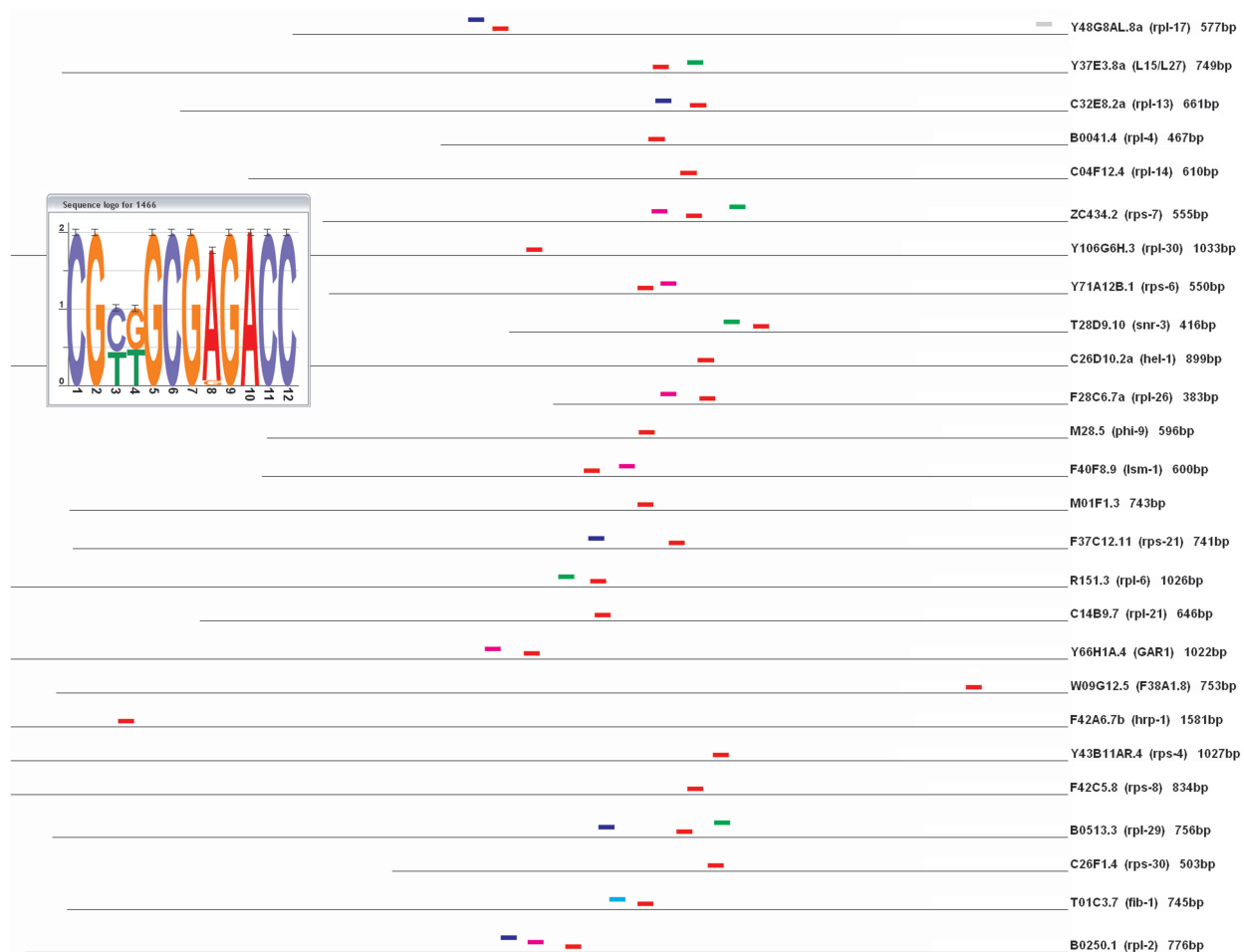


Figure 1. Ribosomal instances of the motif group 12-0. The motif group 12-0 was found upstream of 28 ribosomal transcripts, of which two pairs were on bidirectional promoters: *lsm-1* (*F40F8.9*; a small nuclear ribonucleoprotein splicing factor) and *rps-9* (*F40F8.10*) and *rps-30* (*C26F1.4*) and *rpl-39* (*C26F1.9*). Shown here are the 26 ribosomal upstream regions; instances of motif group 12-0 are shown in red. Instances of motif groups 12-1, 12-5, 12-8, 12-11 and 12-18 are shown in cyan, magenta, gray, blue and green, respectively. The motif logo for all instances of motif group 12-0 in these regions is also shown.

mitochondrial ribosomal genes or both. Mitochondrial ribosomal genes are not specifically annotated in the *C. elegans* genome, but some of them are tentatively identified with KOG (eukaryotic clusters of orthologous groups) designations (39). Of all ribosomal transcripts in the cisRED database, 102 were annotated as ribosomal by KOG, and 24 of these were annotated as mitochondrial ribosomal proteins. Of the 63 ribosomal genes with a ribosomal motif in their upstream region, 50 were further annotated as ribosomal by KOG, and only three of these were also annotated as mitochondrial ribosomal (*B0303.15*, *K07A12.7* and *Y48C3A.10*). The two-tailed *P*-value for this distribution is $4.2E-5$ by the Fisher's exact test; therefore, the motifs we found are specifically associated with cytoplasmic ribosomal genes and not with mitochondrial ribosomal genes.

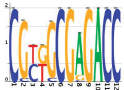
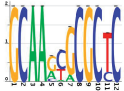
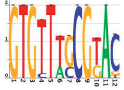
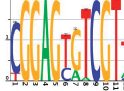
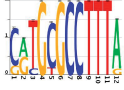
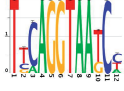
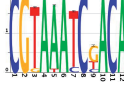

Motifs overlapping trans-splice sites

Motif group 12-8 is an extension of a trans-splice acceptor site. We observed that one of the motifs (12-8) had a

strong strand bias: 28/35 instances were on the same strand as the nearest gene. We also observed that this motif almost always occurs about 20 bp upstream of the ATG. Those instances on the opposite strand tended to be on bidirectional promoters, at the far end of the upstream region, and therefore just upstream of (and on the same strand as) the other gene of the promoter. We also found that 28/35 instances of this motif group overlapped annotated SL1 and SL2 trans-splice sites in Wormbase.

Several other motif groups are extensions of trans-splice acceptor sites. In order to investigate the connection between motif groups and trans-splice acceptor sites more thoroughly, we searched for other motif groups that were also associated with trans-splice acceptor sites (Table 5). Of all 3265 motif groups, at least 16 had a majority of motifs that overlapped with trans-splice sites. Ten of these also had significant associations with ribosomal genes (some were too small in number to have significant associations). All were variations or extensions of the canonical trans-splice site TTTCAG (40).

Table 4. Summary of ribosomal protein-associated motif groups

Group ID	Group name	Background count	Num motifs	Num genes	Num ribosomal	Benjamini <i>P</i> -value	Logo	Characteristics
1466	12-0	200	147	120	28	2.60E-24		Tested via GFP constructs
1467	12-1	162	113	103	8	2.80E-08		Similar to ZF5 site
1469	12-3	118	87	76	10	2.40E-11		Similar to HNF3 family TFBS and EBP-45 site
1470	12-4	86	69	65	6	6.40E-07		
1471	12-5	99	74	65	14	2.80E-16		
1474	12-8	36	35	23	8	1.20E-10		Strand bias; trans-splice site; similar to Delta EF1 site
1477	12-11	123	78	63	12	5.70E-14		
1484	12-18	31	28	21	7	9.30E-04		Similar to <i>Drosophila</i> Zeste site

The first column shows the Group ID of each motif group in the cisRED database, and the second column shows the group name, which also indicates the iteration number of the dodecameric series of motif groups. 'Background Count' shows the number of instances of the motif group sequences among all cisRED upstream regions, and 'Num Motifs' shows the number of instances of the motif group among cisRED motifs. 'Num Genes' shows the number of different genes associated with each motif group, and 'Num Ribosomal' shows how many of these genes were annotated as ribosomal by DAVID, while 'Benjamini *P*-value' indicates the Benjamini-corrected *P*-value of this proportion of ribosomal genes. The logo of each motif group (from all instances, not only ribosomal instances) and other characteristics of each motif group are shown.

GFP testing of function of motif group 12-0

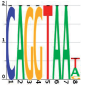
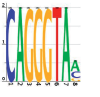
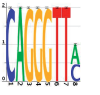
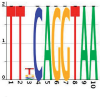
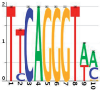
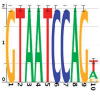
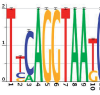
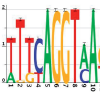
Motif group 12-0 was the largest and the most significant with respect to ribosomal genes, so we investigated further to try to determine its function. We performed a series of GFP experiments to determine whether the presence of the motif was related to gene expression. The BC *C. elegans* Gene Expression Consortium had previously created several thousand GFP constructs and recorded their subsequent expression pattern (26,28). The constructs from the Consortium were made using 3 kb upstream regions or the intergenic region if it was <3 kb. The focus of the Consortium was on genes with human orthologues, but very few ribosomal genes were included. Of the 120 genes with an instance of motif group 12-0 in their upstream regions, 11 had had GFP constructs made by the Consortium, only one of which was a ribosomal gene. However, all 11 of these constructs drove strong expression across a number of different tissues and stages of development, and were therefore good candidates for

further dissection of their promoter regions for assessment of promoter activity.

For each of the 11 upstream regions that had both an instance of motif group 12-0 and previous GFP expression data, three GFP constructs were made: one that included the motif, one that excluded the motif and one that introduced a mutation in the center of the motif (Figure 2). These constructs were injected into the gonad of gravid hermaphrodites, and the worm progeny were allowed to grow to adulthood. Photographs were taken of the worms and their GFP expression was observed and recorded.

GFP expression constructs indicated that motif group 12-0 is involved in regulation of pharyngeal expression. For four of the genes, we found that the construct including the motif produced some GFP expression in the pharynx, while the construct excluding the motif and the construct with a mutated motif showed little or no pharyngeal expression (Table 6 'Positives', Table 7). Two of the

Table 5. Selection of motif groups that overlap trans-splice acceptor sites

Group ID	Group name	Background count	Num motifs	Num trans-splice sites	Num genes	Num ribosomal	Benjamini <i>P</i> -value	Logo
87	8-10	339	224	123	154	19	1.7E-12	
111	8-34	110	76	55	49	11	5.7E-14	
365	8-288	87	47	24	37	9	5.6E-5	
569	10-3	161	119	101	78	21	5.3E-21	
580	10-14	89	66	58	41	13	4.3E-18	
1082	10-516	18	14	11	9	3	6.4E-2	
1474	12-8	36	35	28	23	8	1.2E-10	
2376	14-10	23	22	14	15	4	4.8E-3	

The first column shows the Group ID of the motif group in the cisRED database; the second column shows the group name, which simultaneously indicates the group width and iteration number. 'Background Count' shows the number of instances of the motif group sequences among all cisRED upstream regions, and 'Num Motifs' shows the number of instances of the motif group among cisRED motifs. 'Num Trans-Splice Sites' shows how many of the motifs overlap trans-splice acceptor sites in WormBase; 'Num Genes' shows the number of different genes associated with each motif group, and 'Num Ribosomal' shows how many of these were annotated as ribosomal by DAVID, while 'Benjamini *P*-value' indicates the Benjamini-corrected *P*-value of this proportion of ribosomal genes. The logo of each motif group (from all instances, not only ribosomal instances) is shown.



Figure 2. Schematic of GFP constructs. For each gene with both previous GFP constructs and an instance of motif group 12-0 in its upstream region, three constructs were made. The first construct consisted of the gene's upstream region up to and including the motif but no further (primers indicated by yellow and purple arrows), the second construct was slightly shorter such that the motif was excluded (primers indicated by cyan and purple arrows) and, for the third construct, we introduced a mutation in the central CG of the motif via a primer (primers indicated by orange and purple arrows). Results of the GFP expression assays are shown in Table 6.

genes had inconsistent results because the GFP expression was not correlated with the presence of the motif in the construct (Table 8). Two genes showed no difference in gene expression between the three constructs and were therefore construed as negative results with respect to motif function. Three genes had no GFP expression at all from any of the three constructs, and therefore the function of the motif could not be determined.

DISCUSSION

The DME iterative process, while computationally inefficient, was very effective in identifying sequences that were conserved in the upstream regions of *C. elegans* protein-coding transcripts more often than expected (Table 3). We observed that it tended to skew toward relatively GC-rich sequences because the AT content was considerably lower among cisRED motifs (60.7%) than among the upstream regions (65.8%). It also found the

Table 6. Summary of observed GFP expression

Gene name	Orig. GFP construct		Construct incl. motif		Construct excl. motif		Construct w/mut. motif	
	Pharynx	Other	Pharynx	Other	Pharynx	Other	Pharynx	Other
Expression: Motif:	+		+		-		Mutated	
Positives								
<i>C34E10.6</i>	+++	+++	+	+	-	-	-	-
<i>F09B9.3</i>	+++	+++	+	-	-	-	-	-
<i>F25H2.5</i>	+++	+++	+++	+	+	+	-	+
<i>F54D8.2</i>	+++	+++	+++	++	+	+	-	+
Inconsistent								
<i>C26D10.2</i>	+++	+++	++	++	+	+	-	-
<i>Y57G11C.13</i>	+++	+++	++	++	-	-	+	+
Negatives								
<i>F07A11.2a</i>	+++	+++	++	++	++	++	++	++
<i>T05H4.1</i>	+++	+++	++	++	++	++	++	++
No expression								
<i>C13B9.3</i>	+++	+++	-	-	-	-	-	-
<i>M01F1.3</i>	+++	+++	-	-	-	-	-	-
<i>Y48G8AL.8a</i>	+++	+++	-	-	-	-	-	-

GFP expression is described for four GFP constructs for each of the 11 genes tested in this study: the expression observed by the BC *C. elegans* Gene Expression Consortium ('Orig. GFP Construct'), from the first construct ('Construct Incl. Motif'), from the second construct ('Construct Excl. Motif') and from the third construct ('Construct w/Mut. Motif'). GFP expression is separated into pharyngeal expression and expression in all other tissues because pharyngeal expression showed the greatest differences. The level of GFP expression is indicated by one to three '+', while no GFP expression is indicated by '-'. Genes are sorted into four categories: those that showed a clear difference in expression that correlated with the presence of the motif ('Positives'), those that showed a difference in expression that was not correlated with the presence of the motif ('Inconsistent'), those that showed no difference in expression between the three constructs ('Negatives') and those that showed no GFP expression from any of the three constructs ('No Expression').

largest group first; there was a general trend toward smaller and smaller groups as the DME iterations progressed.

DME counted each instance of a motif group including overlapping sequence instances; palindromes were counted twice. This meant that DME skewed slightly toward repeating and palindromic sequences: the total counts were higher and therefore the difference between the foreground count and the expected foreground count was higher. We did not consider this to be a confounding factor because TFBSs are sometimes palindromic or partially palindromic due to the binding of homodimeric TFs. For example, the *C. elegans* X-box TF DAF-19 binds an imperfect palindromic sequence (41).

A confounding issue was that many of the cisRED motifs overlapped substantially. In a few cases, a series of overlapping cisRED motifs caused DME to identify a sequence as overrepresented when most instances in the foreground referred to a single genomic location. However, some of the upstream regions overlapped as well—bidirectional promoters and alternative transcripts of the same gene—which mitigated the effect of overlapping cisRED motifs somewhat.

We used DAVID to analyze the functional similarities of the genes associated with the first 20 motif groups at each width. An advantage that DAVID has over other GO analysis tools is that it is able to determine whether gene groups are enriched for terms from other gene annotation sources such as the PIR and the KEGG Pathway Database (42) in addition to the GO itself. We found that PIR keywords tended to be both more specific than GO terms and had annotations for more of the genes associated with motif groups, and, as a result, we

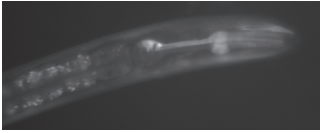
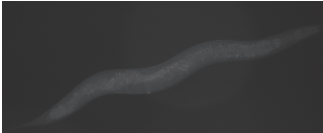

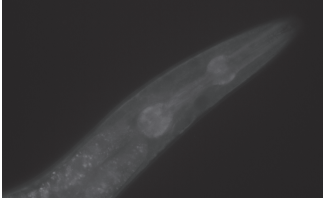
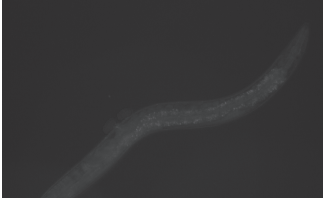
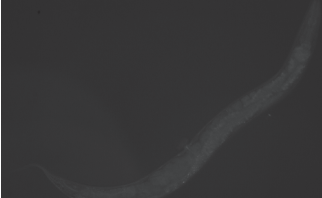
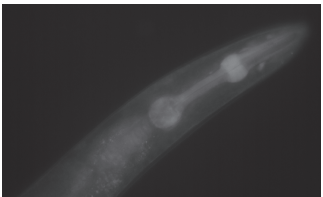
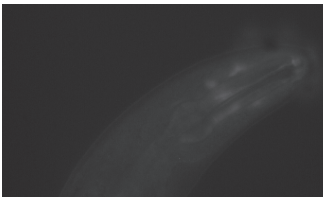
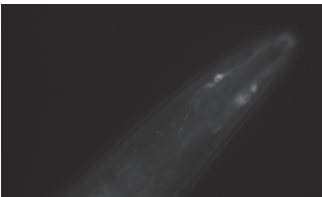
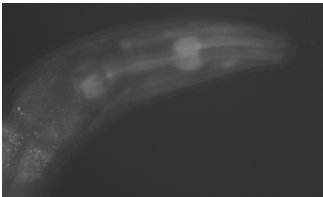
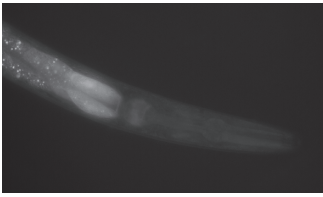
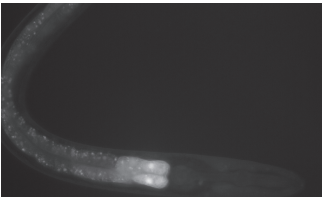
obtained more information about the motif groups than we would have from looking at only GO terms.

We also found that DAVID has several disadvantages. It is a web-based tool that is not designed to be used in a high-throughput way. The HTML-based API is limited both by the maximum URL length and by the internal limit of 400 genes—several of our gene groups exceeded this limit and were not analyzed completely by DAVID. Although it is possible to upload a background gene list for a single gene list, it is not possible to use the correct background list in the API, with the result that some of our significant *P*-values may be off by several orders of magnitude. However, because the *P*-values are extremely low, it is not expected that this incongruity impacted the true significance of any of the motif groups described here.

We used Ontologizer to analyze the functional similarities of genes associated with all motif groups. We found that 19% of the motif groups were associated with genes with significant functional similarities at a *P*-value threshold of 0.05 after one level of multiple testing correction. This result shows that the identification of sequences that are conserved in upstream regions more often than expected—dual analyses of phylogenetic motif discovery followed by motif grouping—is a valid way to find new sequences of interest in a genome. Only a few of the interesting motif groups are discussed here but many others could be investigated in the future.

The advantages of using Ontologizer for this analysis are that, unlike DAVID, it is a true high-throughput analysis program and was able to analyze all 3265 gene groups in a matter of minutes. It was able to ascertain all motif groups with significantly overrepresented GO terms, and as a result we learned that a ninth motif group in the

Table 7. GFP images for positives

Gene	Construct Including Motif	Construct Excluding Motif	Construct with Mutated Motif
<i>C34E10.6</i>			
<i>F09B9.3</i>			
<i>F25H2.5</i>			
<i>F54D8.2</i>			

GFP images for the four upstream regions that resulted in a positive indication of motif function. For each upstream region, the construct including the motif produced GFP expression in the pharynx, while the constructs excluding the motif and with a mutated motif produced no pharyngeal expression.

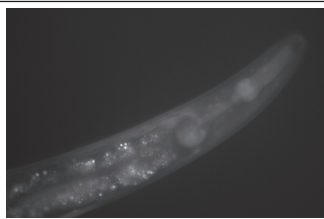
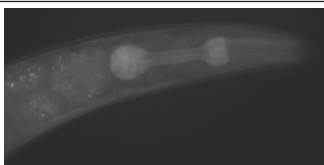
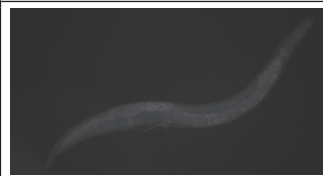
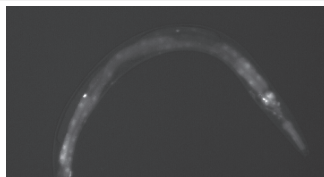
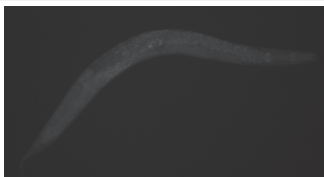
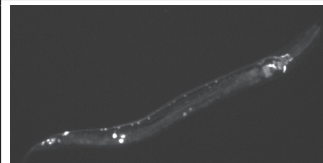
dodecameric series, group 12-31 (group ID 1497) is also associated with ribosomal protein genes even though it is out of the range of the initial analysis by DAVID. It can also use an appropriate background gene list for each GO term test and as a result provided more accurate statistics. The primary disadvantage of Ontologizer is that it relies entirely on the GO for gene annotations and does not refer to other sources such as the PIR and as a result is less sensitive. This difference in annotation sources was highlighted by the disparity in the *P*-values assigned to motif group 12-3: DAVID recorded a minimum *P*-value of 2.4E-11 with 11 of 76 genes annotated as 'ribonucleoprotein' by the PIR, while Ontologizer recorded a minimum *P*-value of 0.04 with 32 of 63 genes annotated as 'embryonic development' by the GO; only six of the genes were annotated as ribonucleoprotein by the GO (*P*-value 0.27).

Any gene group enrichment analysis method will produce some false negatives. A lack of significant

associations does not mean that the genes have nothing further in common. The possibility of gene group significance decreases as the group size decreases; small gene groups (<10 genes) will not be significant unless all of them fall into a specific and rare category. Large gene groups will be highly significant even when only a minority of the genes fall into the same category.

For hexameric and octameric motif groups, DME terminated automatically after 76 and 489 iterations, respectively, with no further motif groups left to find. For decameric, dodecameric, and width-14 motif groups, we terminated DME after 900 iterations because the low information content requirement made it possible for DME to find a virtually unlimited number of very small motif groups. Of the 3265 motif groups, 2775 (85%) consisted of sequence patterns that were significantly overrepresented among cisRED motifs with respect to the upstream regions with a *P*-value <0.05 as calculated by the binomial test. Motif group significance tailed off

Table 8. GFP images for inconsistent results

Gene	Construct Including Motif	Construct Excluding Motif	Construct with Mutated Motif
<i>C26D10.2</i>			
<i>Y57G11C.13</i>			

GFP images for the two upstream regions that resulted in an inconsistent indication of motif function. For each upstream, the construct including the motif produced GFP expression in the pharynx. For *C26D10.2*, the construct excluding the motif also produced some pharyngeal expression and the construct with a mutated motif produced no expression. For *Y57G11C.13*, the construct excluding the motif produced no expression and the construct with a mutated motif produced GFP expression in a variety of tissues.

slowly as the iterations progressed. However, only 612 (19%) of the motif groups were associated with genes that also had significant functional similarities with respect to the GO as calculated by Ontologizer. For widths 10, 12 and 14 bp, the distribution of Ontologizer *P*-values showed a sharp peak of highly significant *P*-values for the first 50 iterations, followed by a long flat tail of insignificant *P*-values punctuated by occasional marginally significant groups. Because of this distribution, we expect that 900 iterations were sufficient to find most groups of interest to future research.

Eight of the top 20 motif groups in both decameric and dodecameric series of DME iterations were enriched for ribosomal genes. We decided to concentrate our efforts on the dodecameric motif groups because, although both sets had similar *P*-values and were probably equally valid, the dodecameric groups had fewer genes associated with them, making further analysis more straightforward.

Several of the ribosomal motif groups were similar to experimentally validated TFBSs as determined by TESS (Table 4). Motif group 12-1 was found to be similar to mouse ZF5, but the section of the motif group that matched ZF5 was also the least-conserved portion, so it is unlikely that the similarity is important; additionally, *C. elegans* has no known ZF5 ortholog. The similarity of motif group 12-3 to EBP-45 and HNF3-family TFs, which bind to similar sequences even though they are structurally unrelated, may be more interesting. *Caenorhabditis elegans* has two genes that are putative EBP orthologues, *zip-4* and *cebp-1*, and two genes that encode HNF3-family TFs: *lin-31* and *pha-4* (43). The similarity of motif group 12-8 to the rat Delta EF1 site is probably not important. The site sequence is not a perfect match, and it seems likely that the importance of this site and the other trans-splice acceptor site-related motif groups is due to splicing factors binding to the RNA, not TFs binding to the DNA. The similarity of

motif group 12-18 to the *Drosophila* Zeste site is worthy of note: the site similarity is in the perfectly conserved portion of the motif group. Additionally, Zeste is a polycomb-group protein that has a known ortholog in *C. elegans*: MES-2 (44). The *C. elegans mes-2* knockout mutant has the maternal effect sterile phenotype, which means that MES-2 is an important protein required for germline development.

Closer inspection of the strand- and location-biased motif group 12-8 revealed that not only was it an extension of a trans-splice acceptor site, but also that several other motif groups were also trans-splice acceptor site extensions (Table 5). It makes sense that the trans-splice locations would be conserved in the orthologous regions, as it is logical that the other nematodes also perform trans-splicing of transcripts. The canonical trans-splice site is TTTCAG; our results suggest that the trans-splice acceptor site may in fact be more complex. One of the motif groups (10-516) was a noncanonical extension at the 5' end of the trans-splice site: for nine genes (of which three were confirmed ribosomal), we saw the pattern GTAATCCAG at the trans-splice site. The other motif groups were all extensions of the trans-splice site at the 3' end, beyond the CAG. There were three specific extensions of the pattern: CAGGTAA (motif groups 8-10, 10-3, 12-8 and 14-10), CAGGGTA (motif groups 8-34 and part of 10-14) and CAGGGTT (motif groups 8-288 and part of 10-14).

Graber *et al.* (45) identified a series of pentamers that are overrepresented immediately after the splice junction of SL1 trans-splice acceptor sites with respect to the same location in SL2 acceptor sites and intronic splice sites. Five of the pentamers described in this study are similar to the motif groups shown in Table 5 as follows: pentamer GGGUA was overrepresented in SL1 sites with a *P*-value of 1.5E-5 and is similar to motif groups 8-34 and 10-14; pentamer GGGUU (*P*-value 2.0E-5) is similar to motif

groups 8-228 and 10-14; pentamer GGUAA (P -value $4.9E-6$) is similar to motif groups 8-10, 10-3, 12-8 and 14-10; pentamer GUAAA (P -value $2.9E-5$) and pentamer GUAAU (P -value $9.8E-6$) are both similar to motif group 8-10. The similarities between the results of that study and this one provide further evidence that the patterns described here are important. However, it is not clear why ribosomal genes in particular would have such variant trans-splice acceptor sites.

Although it is clear that the ribosomal genes discussed here are in general enriched for cytoplasmic ribosomal genes, it is possible that one or more of the motif groups is associated with mitochondrial ribosomal genes. The three KOG-annotated mitochondrial ribosomal genes only had instances of motif groups 12-3 and 12-4 in their upstream regions. These same two motif groups were found to co-occur the least with other ribosomal motif groups. In contrast, half of the instances of motif group 12-0 were found to occur in close proximity to instances of motif groups 12-5, 12-11 or 12-18 in a striking pattern (Figure 1). Without more information as to the specific function of each of these ribosomal genes, it is difficult to investigate these occurrence patterns in more depth.

Ribosomal proteins have been shown to be tightly regulated (46). Most of the proteins must be produced at equimolar levels in order for ribosomes to be synthesized correctly, and several ribosomal proteins regulate the splicing of their own mRNA. Incorrect proportions of ribosomal proteins and subsequent partially synthesized ribosomal particles can lead to cell cycle arrest and apoptosis in vertebrates (46).

Four of the 11 genes tested using GFP constructs displayed a dependence on motif group 12-0 for pharyngeal expression (Tables 6 and 7). It is not clear why the motif was related to expression in the pharynx as opposed to other tissues, because these genes are normally expressed in most or all tissues. For one of the positive results (*F09B9.3*), the motif was in the WormBase-annotated 5' UTR, suggesting that the genome contains an additional transcription start site between the motif and the gene's ATG. The motif seemed to be better-conserved in the upstream regions of the genes that had positive indications of function, but most of them were well enough conserved to be found repeatedly by motif discovery of the upstream regions and their orthologs. The motif was very poorly conserved for two of the 11 genes (*Y57G11C.13* and *T05H4.1*, which had 'inconsistent' and 'negative' results, respectively). They were so poorly preserved that although they were found within cisRED motifs in an earlier unpublished version of the cisRED database, they were not within motifs in the published cisRED database. Only one ribosomal gene was tested with GFP (*rpl-17*, or *Y48G8AL.8a*), but because no GFP expression was observed for any of the three constructs, we were unable to determine whether the motif was involved in the regulation of the expression of this gene. It is important to note that the GFP assay relied on the assumption that there were no important elements upstream of the tested motif that affected the transcription of the genes in question, because the largest construct only extended up to the

end of the motif being tested and no further. This assumption may not have been true for some of the genes. The existence of other elements further upstream may explain why no GFP expression was seen for three of the tested genes.

Similarly, the assay also relied on the assumption that there were no important elements between the 'primer including motif' and the 'primer excluding motif'. We designed the two primers to be as close together as possible but were limited by melting temperature considerations. An alternate interpretation of the 'inconsistent' expression pattern result for *Y57G11C.13* is that an element located between the two primers was responsible for the different expression levels and the identified motif was not involved in the reporter gene expression.

CONCLUSIONS

The motif grouping program DME was successful in finding interesting sequences that were conserved in the orthologs much more often than expected. The motif groups had significant functional associations, showing that the repeated, evolutionarily conserved sequences that we found could not have occurred by chance and have biological importance. The P -values for the ribosomal motif groups were extremely low after multiple testing correction was performed, and robust in the sense that similar statistics were calculated repeatedly, regardless of variations in the width of the motif and the IC.

Two of the eight ribosomal motifs are similar to known binding sites of TFs that have *C. elegans* orthologs and warrant further investigation of this connection. Trans-splice sites are strongly conserved for ribosomal genes and follow specific patterns that are extensions of the canonical trans-splice sites. Motif group 12-0 is usually found 300bp upstream of the ATG of ribosomal genes and tends to occur in close proximity to instances of motif groups 12-5, 12-11 or 12-18. GFP construct experiments in broadly expressed genes indicated that it may have a direct impact on the pharyngeal expression of some genes, but its influence on the expression of ribosomal genes remains undetermined.

ACKNOWLEDGEMENTS

The authors are grateful to Andrew D. Smith for providing a custom-built version of DME.

FUNDING

This research was funded in part by Genome Canada, Genome British Columbia and the British Columbia Cancer Foundation. MCS is a Michael Smith Foundation for Health Research Trainee. AKM is supported by a Natural Sciences and Engineering Research Council of Canada PGS-D Graduate Scholarship. DLB is a Canada Research Chair in Genomics and is funded by a grant from the Natural Sciences and Engineering Research Council of Canada. SJMJ is a Michael Smith Foundation for Health

Research Senior Scholar. Funding for open access charge: Genome Canada.

Conflict of interest statement. None declared.

REFERENCES

- Sleumer, M.C., Bilenky, M., He, A., Robertson, G., Thiessen, N. and Jones, S.J. (2009) *Caenorhabditis elegans* cisRED: a catalogue of conserved genomic elements. *Nucleic Acids Res.*, **37**, 1323–1334.
- Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P. and Moreau, Y. (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.*, **9**, 447–464.
- Griffith, O.L., Montgomery, S.B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M.C., Bilenky, M., Haeussler, M. et al. (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.*, **36**, D107–D113.
- Bryne, J.C., Valen, E., Tang, M.E., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B. and Sandelin, A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, 102–106.
- Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K. et al. (2006) TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, 108–110.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
- He, H., Wang, J., Liu, T., Liu, X.S., Li, T., Wang, Y., Qian, Z., Zheng, H., Zhu, X., Wu, T. et al. (2007) Mapping the *C. elegans* noncoding transcriptome with a whole-genome tiling microarray. *Genome Res.*, **17**, 1471–1477.
- Stumpf, C.R., Kimble, J. and Wickens, M. (2008) A *Caenorhabditis elegans* PUF protein family with distinct RNA binding specificity. *RNA*, **14**, 1550–1557.
- de Wit, E., Linsen, S.E., Cuppen, E. and Berezikov, E. (2009) Repertoire and evolution of miRNA genes in four divergent nematode species. *Genome Res.*, **19**, 2064–2074.
- Roush, S. and Slack, F.J. (2008) The let-7 family of microRNAs. *Trends Cell Biol.*, **18**, 505–516.
- Lin, Y.C., Hsieh, L.C., Kuo, M.W., Yu, J., Kuo, H.H., Lo, W.L., Lin, R.J., Yu, A.L. and Li, W.H. (2007) Human TRIM71 and its nematode homologue are targets of let-7 microRNA and its zebrafish orthologue is essential for development. *Mol. Biol. Evol.*, **24**, 2525–2534.
- Wang, T. and Stormo, G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.
- Siddharthan, R., Siggia, E.D. and van Nimwegen, E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, **1**, e67.
- Zhao, G., Schriefer, L.A. and Stormo, G.D. (2007) Identification of muscle-specific regulatory modules in *Caenorhabditis elegans*. *Genome Res.*, **17**, 348–357.
- Van Loo, P. and Marynen, P. (2009) Computational methods for the detection of cis-regulatory modules. *Brief Bioinform.*, **10**, 509–524.
- Smith, A.D., Sumazin, P. and Zhang, M.Q. (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc. Natl Acad. Sci. USA*, **102**, 1560–1565.
- Smith, A.D., Sumazin, P., Xuan, Z. and Zhang, M.Q. (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc. Natl Acad. Sci. USA*, **103**, 6275–6280.
- Smith, A.D., Sumazin, P. and Zhang, M.Q. (2007) Tissue-specific regulatory elements in mammalian promoters. *Mol. Syst. Biol.*, **3**, 73.
- Minovitsky, S., Stegmaier, P., Kel, A., Kondrashov, A.S. and Dubchak, I. (2007) Short sequence motifs, overrepresented in mammalian conserved non-coding sequences. *BMC Genomics*, **8**, 378.
- Das, D., Clark, T.A., Schweitzer, A., Yamamoto, M., Marr, H., Arribere, J., Minovitsky, S., Poliakov, A., Dubchak, I., Blume, J.E. et al. (2007) A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. *Nucleic Acids Res.*, **35**, 4845–4857.
- Eguchi, J., Yan, Q.W., Schones, D.E., Kamal, M., Hsu, C.H., Zhang, M.Q., Crawford, G.E. and Rosen, E.D. (2008) Interferon regulatory factors are transcriptional regulators of adipogenesis. *Cell. Metab.*, **7**, 86–94.
- Barrera, L.O., Li, Z., Smith, A.D., Arden, K.C., Cavenee, W.K., Zhang, M.Q., Green, R.D. and Ren, B. (2008) Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. *Genome Res.*, **18**, 46–59.
- Benita, Y., Kikuchi, H., Smith, A.D., Zhang, M.Q., Chung, D.C. and Xavier, R.J. (2009) An integrative genomics approach identifies hypoxia inducible factor-1 (HIF-1)-target genes that form the core response to hypoxia. *Nucleic Acids Res.*, **37**, 4587–4602.
- Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
- Bauer, S., Grossmann, S., Vingron, M. and Robinson, P.N. (2008) Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, **24**, 1650–1651.
- Hunt-Newbury, R., Viveiros, R., Johnsen, R., Mah, A., Anastas, D., Fang, L., Halfnight, E., Lee, D., Lin, J., Lorch, A. et al. (2007) High-throughput in vivo analysis of gene expression in *Caenorhabditis elegans*. *PLoS Biol.*, **5**, e237.
- Hobert, O. (2002) PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic *C. elegans*. *BioTechniques*, **32**, 728–730.
- McKay, S.J., Johnsen, R., Khattri, J., Asano, J., Baillie, D.L., Chan, S., Dube, N., Fang, L., Goszczynski, B., Ha, E. et al. (2003) Gene expression profiling of cells, tissues, and developmental stages of the nematode *C. elegans*. *Cold Spring Harb. Symp. Quant. Biol.*, **68**, 159–169.
- Thacker, C., Sheps, J.A. and Rose, A.M. (2006) *Caenorhabditis elegans* dpy-5 is a cuticle procollagen processed by a proprotein convertase. *Cell Mol. Life Sci.*, **63**, 1193–1204.
- Wu, C.H., Yeh, L.S., Huang, H., Arminski, L., Castro-Alvarez, J., Chan, Y., Hu, Z., Kourtesis, P., Ledley, R.S., Suzek, B.E. et al. (2003) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.
- Schug, J. (2003) Using TESS to predict transcription factor binding sites in DNA sequence. In Baxevanis, A.D. (ed.), *Current Protocols in Bioinformatics*. J. Wiley and Sons.
- Obata, T., Yanagidani, A., Yokoro, K., Numoto, M. and Yamamoto, S. (1999) Analysis of the consensus binding sequence and the DNA-binding domain of ZF5. *Biochem. Biophys. Res. Commun.*, **255**, 528–534.
- Petropoulos, I., Auge-Gouillou, C. and Zakin, M.M. (1991) Characterization of the active part of the human transferrin gene enhancer and purification of two liver nuclear factors interacting with the TGTTTGC motif present in this region. *J. Biol. Chem.*, **266**, 24220–24225.
- Grange, T., Roux, J., Rigaud, G. and Pictet, R. (1991) Cell-type specific activity of two glucocorticoid responsive units of rat tyrosine aminotransferase gene is associated with multiple binding sites for C/EBP and a novel liver-specific nuclear factor. *Nucleic Acids Res.*, **19**, 131–139.
- Cardinaux, J.R., Chapel, S. and Wahli, W. (1994) Complex organization of CTF/NF-I, C/EBP, and HNF3 binding sites within the promoter of the liver-specific vitellogenin gene. *J. Biol. Chem.*, **269**, 32947–32956.
- Sekido, R., Murai, K., Funahashi, J., Kamachi, Y., Fujisawa-Sehara, A., Nabeshima, Y. and Kondoh, H. (1994) The delta-crystallin enhancer-binding protein delta EF1 is a repressor of E2-box-mediated gene activation. *Mol. Cell. Biol.*, **14**, 5692–5700.

37. Benson, M. and Pirrotta, V. (1988) The *Drosophila* zeste protein binds cooperatively to sites in many gene regulatory regions: implications for transvection and gene regulation. *EMBO J.*, **7**, 3907–3915.
38. Suzuki, T., Terasaki, M., Takemoto-Hori, C., Hanada, T., Ueda, T., Wada, A. and Watanabe, K. (2001) Proteomic analysis of the mammalian mitochondrial ribosome. Identification of protein components in the 28 S small subunit. *J. Biol. Chem.*, **276**, 33181–33195.
39. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
40. Fields, C. (1990) Information content of *Caenorhabditis elegans* splice site sequences varies with intron length. *Nucleic Acids Res.*, **18**, 1509–1512.
41. Efimenko, E., Bubb, K., Mak, H.Y., Holzman, T., Leroux, M.R., Ruvkun, G., Thomas, J.H. and Swoboda, P. (2005) Analysis of *xbx* genes in *C. elegans*. *Development*, **132**, 1923–1934.
42. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
43. Bieri, T., Blasiar, D., Ozersky, P., Antoshechkin, I., Bastiani, C., Canaran, P., Chan, J., Chen, N., Chen, W.J., Davis, P. *et al.* (2007) WormBase: new content and better access. *Nucleic Acids Res.*, **35**, D506–D510.
44. Holdeman, R., Nehrt, S. and Strome, S. (1998) MES-2, a maternal protein essential for viability of the germline in *Caenorhabditis elegans*, is homologous to a *Drosophila* Polycomb group protein. *Development*, **125**, 2457–2467.
45. Graber, J.H., Salisbury, J., Hutchins, L.N. and Blumenthal, T. (2007) *C. elegans* sequences that control trans-splicing and operon pre-mRNA processing. *RNA*, **13**, 1409–1426.
46. Warner, J.R. and McIntosh, K.B. (2009) How common are extraribosomal functions of ribosomal proteins? *Mol. Cell*, **34**, 3–11.