Three Genes, Two Species: A Comparative Analysis of Upstream Regulatory Sequences Sufficient to Direct Vulval Expression in *C. elegans* and *C. briggsae*

Martha Kirouac and Paul W. Sternberg

(submitted for publication)

ABSTRACT

We have identified the *Caenorhabditis briggsae* homologs of three *C. elegans* genes, *egl-17*, *zmp-1* and *cdh-3*, that are differentially expressed in subsets of vulval cells and the anchor cell. Upstream cis-regulatory regions of the *C. elegans* genes sufficient to confer vulval and anchor cell specific regulation are known (Kirouac and Sternberg, accompanying manuscript). We have identified the corresponding *C. briggsae* control regions and tested these regions for activity in *C. elegans.* We find that a 748-bp region of *C. briggsae egl-17* confers expression in *C. elegans* in the primary lineage, occasional secondary lineage expression and late expression in vulC and D. We have identified a 755-bp upstream region of *C. briggsae zmp-1* that confers expression in vulE, vulA, and the anchor cell in *C. elegans*. Finally, we have identified a 1.4-kb region of *C. briggsae cdh-3* that drives expression in vulE, F, C, and D cells in *C. elegans,* and a separate 277-bp region of *C. briggsae cdh-3* that confers expression to *C. elegans* vulC, E and F, but not vulD. We conclude that these phylogenetic footprints promote vulval cell expression in both species. Lastly, we compare the efficacy of phylogenetic footprinting with respect to deletion analysis in transgenic animals.

INTRODUCTION

One of the hallmarks of metazoan development is the transition of an undifferentitated population of cells into unique terminal-cell types. Intercellular signaling plays a major role in the differentiation of cell populations compared to the number of cell types, but, there are relatively few signaling pathways that specify a broad range of terminal fates. The mechanisms by which unique populations of cells are generated from these general signaling components are not well understood.

In the development of the *C. elegans* vulva, at least three intercellular signaling pathways, the EGF, NOTCH, and WNT pathways, induce six multipotential Vulval Precursor Cells (VPCs; reviewed in Greenwald, 1997; Kornfeld, 1997; Sternberg and Han, 1998) to generate an invariant spatial pattern of seven cell fates; vulA-F (Sharma-Kishore *et al*., 1999). This patterning is likely to depend upon the cis-regulatory regions of the transcriptional targets of these intercellular signals. The isolation of response elements in their transcriptional targets will facilitate biochemical and bioinformatic identification of major transcriptional factors that control cell specific gene expression downstream of these canonical signaling pathways.

Regulatory regions sufficient for vulva and anchor cell expression of three target genes have been described (Kirouac and Sternberg, in prep.): *egl-17*, a fibroblast growth factor family member; *zmp-1*, which encodes a zinc metalloproteinase gene; and *cdh-3*, which encodes a FAT-like cadherin gene. These sufficiency regions probably encode multiple binding sites spread over an extended area. To delimit what regions might be the most important in determining vulva and anchor cell specificity, we have identified the *C. briggsae* homologs of these three genes, and then used phylogenetic footprinting to

identify the control regions predicted to correspond to the sufficiency regions in *C. elegans*. Phylogenetic footprinting is a method for the identification of regulatory elements in a set of orthologous regulatory regions from multiple species by identifying the best-conserved motifs in those regions (Tagle *et al*., 1988).

Despite having diverged from one another an estimated 50-120 million years ago (Coghlan, 2002), both *C. elegans* and *C. briggsae* share almost identical development and morphology (Nigon and Dougherty, 1949), and the sequences of both species are now known. Rescue of *C. elegans* mutant phenotypes with *C. briggsae* has demonstrated that there is functional conservation between the two species (e.g. de Bono and Hodgkin, 1996; Kennedy *et al*., 1993; Krause *et al*., 1994; Kuwabara, 1996; Maduro and Pilgrim, 1996). In addition, analysis of similarity within 142 pairs of orthologous intergenic regions shows regions of high similiarity interspersed with non-alignable sequence (Webb *et al*., 2002). The high degree of similarity in some of these regions suggests that they are under selective pressure. Such intergenic conservation between *C. elegans* and *C. briggsae* has been utilized in various studies to isolate putative binding sites for trans-acting regulatory factors (e.g., Culetto *et al*., 1999; Gilleard *et al*., 1997; Gower *et al*., 2001; Krause *et al*., 1994; Xue *et al*., 1992).

In this paper, we test intergenic conserved regions from *C. briggsae* for their ability to drive GFP expression in the vulva cells and anchor cell from the basal *pes-10* promoter for expression in both *C. elegans* and *C. briggsae*.

MATERIALS AND METHODS

Protein prediction of EGL-17, ZMP-1, and CDH-3 homologs in *C. briggsae*

The sequence of the *C. elegans* translated protein used for the TBLASTX was obtained either through Wormbase (http://www.wormbase.org/; Stein et al., 2001), as was the case for EGL-17 and CDH-3, or from personal communication in the case of ZMP-1 (J. Butler and J. Kramer, personal communication). For each of these three predicted genes, the corresponding *C. briggsae* cDNA was partially sequenced from an RT-PCR product made from poly $(A)^+$ RNA that was isolated from mixed-staged *C. briggsae* worms. The following primers were used for RT-PCR: mk166 5' AGGCGAAACCCACTGGCAAC 3' and mk167 5' TTTGGCGGAGCAGAACACAC 3' for *egl-17*; mk168 5' ATGGGTATT TGCCCCGTGGC 3' and mk169 5' GATTTCCTTCTCATAGGTGAACGC 3' for *zmp-1*; and mk170 5' CCTCTCCAACTCGACATGAATCTC 3' and mk171 5' ACAGTCAAGT TTTCGATTGCGG 3' for *cdh-3*.

Analysis of homologous upstream sequences in *C. elegans* and *C. briggsae*

The Seqcomp and Family Relations programs (Brown *et al*., 2002) were used to identify homologous upstream sequences conserved between *C. elegans* and *C. briggsae*. The Seqcomp algorithm compares a window of fixed size from one sequence against a same sized window in the second sequence. All 20-bp windows were compared between the two species, at an 80-85% threshold level. This threshold level allows three to four mismatches in a 20-bp window. The upstream sequences of *egl-17*, *zmp-1* and *cdh-3* lie on *C briggsae* contigs c000300114, c010400937, and c01090600, respectively.

Generation of *egl-17, zmp-1* and *cdh-3 C. briggsae* promoter GFP constructs

Using PCR primers designed from the predicted conserved regions between the upstream regions of *C. elegans* and *C. briggsae egl-17, zmp-1* and *cdh-3,* the regions of interest were amplified, with TaKaRa LA Taq (Takara Shuzo), and cloned into the minimal promoter *pes-10,* pPD107.94 (a gift from the Fire lab) using Sph I (5') and Xba I (3') restriction sites engineered into the primers. The sequence of these primers were as follows: mk160, 5' CCCCCGCATGCCACGACCTCCTGGTGTGAGG 3', and mk161, 5' CCCCCTCTAGACTAACAA ATGACAAGCGGAAG 3', for *egl-17*; mk172, 5' CCCCC GCATGCGAGTTTCTGGAG GATTCTG 3', and mk173, 5' CCCCCTCTAGACGGAA TACTTTAGAATCTC 3', for *zmp-1*; mk162, 5' CCCCCGCATGCCTGACTATGGGGC AGGTGGCC 3', and mk163, 5' CCCCCTCTAGAGGTGCGGGAAGAGCCGAGC 3', for the *cdh-3* region containing elements A-F; mk164, 5' CCCCCGCATGCGTCTGTTT GTCCCGATGTCGA 3', and mk165, 5' CCCCCTCTAGAGTAGATGGCTGGGATGA CAGG 3', for the *cdh-3* region containing elements H-K. The following PCR protocol was used: 94.0 °C for 4 minutes, followed by 30 cycles 94.0 °C for 30 seconds, 58.0-60.0°C for 30 seconds, 68.0 °C for 7 minutes, followed by 7 minutes at 68.0 °C. *C. briggsae* genomic DNA served as a template for the PCR reaction.

The nomenclature of the constructs generated in this study is derived from the primers used to amplify the region. In all cases, the first 1-3 digits represent the 5' primer and the digits after the hyphen represent the 3' primer.

Microinjection of promoter GFP constructs into *C. elegans*

The constructs were microinjected into the gonads of animals of genotype *pha-1*(e*2123ts*); *him-5*(e*1490*) line using a standard protocol (Mello et al., 1991). The constructs were injected at a concentration of 100 ng/µl, with 20 ng/µl pBluescript SKII (Stratagene), and 82 ng/µl *pha-1*(+), pBX. Transgenic animals that stably transmit the extrachromosomal arrays were isolated by selecting viable F1 animals at 22.0 °C to new plates and examining their progeny for GFP expression in the anchor cell, and the vulval cells.

Microinjection of promoter GFP constructs into *C. briggsae*

The constructs were microinjected into the gonads of AF16, a wild-type *C. briggsae* line (Fodor *et al*., 1983), using a standard protocol (Mello *et al*., 1991). Constructs were injected at a concentration of 100 ng/µl, with 110 ng/µl pBluescript- SKII, and 10 ng/µl *myo-2*::GFP. Transgenic animals stably transmitting the extra-chromosomal arrays were isolated by selecting for *myo-2*::GFP expression in the pharynx of F2 animals. These animals were transferred to new plates, and lines that stably transmitted the array were examined for vulva GFP expression in their progeny.

Microscopy of transgenic animals

Animals were mounted on 5% noble agar pads and scored at 20.0°C for GFP expression under Nomarski optics using a Zeiss Axioplan microscope with a 200-watt HBO UV source, and a Chroma High Q GFP LP filter set (450 nm excitation/505 nm emission). At least two lines for each construct were examined.

egl-17 early expression in the granddaughters of P6.p, the precursors of vulE and vulF cells, was scored at the four-cell stage. *egl-17* vulC and vulD GFP expression was scored between the late L4 to young adult stages (Burdine *et al*., 1998). *zmp-1* anchor cell GFP expression was scored between the L3 and the early L4 stage. VulE and vulD expression was scored between late L4 and young adult stages. *zmp-1* vulA expression was scored between young adult and adult stages (Wang and Sternberg, 2000). *cdh-3* AC GFP expression was scored between the L3 and the early L4 stage. *cdh-3* vulE, vulF, vulC and vulD expression was scored between the late L3 stage through late L4 stages (Pettitt *et al*., 1996).

Prediction of binding sites using Transfac database

Putative binding sites for known transcription factors in the conserved regions defined by comparative analysis between *C. elegans* and *C. briggsae* in the *egl-17*, *zmp-1* and *cdh-3* upstream regions were determined using the Transfac database and the MatInspector program (http://www.genomatix.de/mat_fam; Quandt *et al*., 1995). Particular emphasis was placed on the regions that were sufficient to confer expression in transgenic *C. elegans* on *pes-10* (Kirouac and Sternberg, in prep.).

AlignACE predictions of overrepresented sequences

AlignACE is based on a Gibbs sampling algorithm that computes a series of motifs that are over-represented in the input sequence(s) (http://atlas.med.harvard.edu/cgi-bin/alignace.pl; Roth *et al.,* 1998). The MAP score (maximum a priori log likelihood) is the functional readout of the degree to which a motif is over-represented relative to the

expectation for the random occurrence of such a motif in the sequence under consideration (Roth *et al.,* 1998). We chose a MAP cut-off of 10, which has been shown to be adequate to identify the best-studied examples of known transcription factor binding sites in yeast (Hughes *et al.*, 2000). We used a GC content setting of 0.35, and we searched for motifs of eight and 10 nucleotides. A greater number of aligned sites that are more tightly conserved with information-rich positions, and with nucleotides that are less prevalent in the genome, will lead to higher MAP scores (Hughes *et al.*, 2000).

RESULTS

C. briggsae homologs of *egl-17, zmp-1* and *cdh-3*

Because genomic regions that have a biological function are often conserved through evolution, non-coding regions conserved between species are more likely to contain regulatory sequences (Stern, 2000). Therefore, we examined *egl-17*, *zmp-1* and *cdh-3* in the related nematode species, *C. briggsae*.

To identify conserved upstream regulatory regions, we first identified the homologs of ZMP-1, EGL-17 and CDH-3 in *C. briggsae.* Predictions of the *C. briggsae* cDNAs were based on TBLASTX searches of Jim Mullikin's PHUSION assembler data (11/11/2001) at Washington University (http://genome.wustl.edu/gsc/), combined with prediction of splice-site donor and acceptor sites using the NetGene2 (http://www.cbs.dtu.dk/services/NetGene2/) program. The *C. briggsae* cDNAs were isolated from mixed-staged poly $(A)^+$ RNA and sequenced using primers based on these predictions.

The predicted *C. briggsae* EGL-17 cDNA lies on contig c000300114. As seen in the ClustalW alignment (Figure 1), the EGL-17 proteins in both species consist of five translated exons. The *C. elegans* protein has 216 amino acids, and the predicted *C. briggsae* protein has 218 amino acids. The *C. briggsae* exons three and four were sequenced, as were most of exons two and five (Genbank accession #AF529234). The six beta strands and three hairpin structural domains that make up the beta trefoil-fold structural element of the FGF ligand family is conserved in this prediction.

The predicted ZMP-1 *C. briggsae* cDNA lies on two non-overlapping contigs, c010400937 and c000100134. As seen in the ClustalW alignment (Figure 2), the *C. elegans* ZMP-1 protein consists of eight translated exons, as does the *C. briggsae* protein. The *C. elegans* protein has 521 amino acids, and the predicted *C. briggsae* protein has 517 amino acids. There are several interesting features of the sequences. First, the length of the large third intron of approximately 3 kb is conserved in both species. Second, the *C. briggsae* genomic sequence has a large intron of ≥5 kb after exon six, where the sequence jumps between non-overlapping contigs. The cDNA from *C. briggsae* was sequenced and the prediction was confirmed for the entirety of exons four, five and six, and most of exons three and seven (Genbank accession #AF529235). Additionally, the conserved matrix metalloproteinase motif, HEXXH, was sequenced and found to be conserved in the sixth exon, and the predicted PRCGXPD motif of the matrix metalloproteinase family located in the second exon is conserved in the prediction.

The predicted *C. briggsae* CDH-3 cDNA lies on two overlapping contigs, c014100642 and c01090600. As seen in the ClustalW alignment (Figure 3), the *C. elegans* protein consists of 23 translated exons, while the *C. briggsae* protein consists of

21 exons. Exons three and four in *C. elegans* are present in a single exon in *C. briggsae*. Similarly, the exons corresponding to *C. elegans* exons nine and ten are present in exon eight in *C. briggsae*, and exons 18 and 19 in the *C. elegans* transcript are represented by exon 16 in *C. briggsae*. Finally, exon 21 from *C. elegans* is split into exons 18 and 19 in *C. briggsae*. Overall, the *C. elegans* protein has 3343 amino acids, and the predicted *C. briggsae* protein has 3221 amino acids. The cDNA from *C. briggsae* was sequenced, and the prediction was confirmed for exons three through five, and parts of exons two and six (Genbank accession #AF529236). The eleven predicted cadherin domains, and the lamin G domain in the *C. elegans* protein (wormPD report CDH-3 at http://www.incyte.com/proteome/WormPD; Costanzo *et al*., 2000) are conserved in the *C. briggsae* prediction.

Comparative sequence analysis

Previous comparisons of intergenic regions have relied on gross alignment of these sequences to find regions of similarity using ClustalW (Higgins *et al.*, 1996) or other alignment programs. In our analysis, we used the Seqcomp and Family Relations programs that perform a comparison of two genomic sequences (Brown *et al*., 2002). This algorithm allows the isolation of possible conserved regions regardless of location or orientation (i.e., this allows the isolation of similarities from the reverse complement of the sequence). Regions of high similarity between two species such as *C. elegans* and *C. briggsae* are termed phylogenetic footprints (Tagle *et al*., 1988). The footprints between these two species are, on average, 80% similar, while whole intergenic regions are, on average, 47% similar in *C. elegans* and 50% similar in *C. briggsae* (Webb *et al*., 2002).

Therefore, a comparison of these regions at a threshold value of 85-90% identity should allow selection of the most similar non-coding regions.

For the *egl-17* comparison, we used the entire 3.9 kb genomic region upstream of the translational start site in *C. elegans* as a basis for comparison against the *C. briggsae* sequence upstream of the predicted *egl-17* translational start site. At the 90% threshold level, four regions of similarity are found (Figure 4A). These elements (A, B, C and D) were located in the same orientation and order with respect to each other in the two species (Figure 5). Elements B, C and D all appear at a 100% threshold level, and at lower thresholds, these regions expand. Element A shares 90% identity between the two species. Two of these four elements, B and D, are in regions of the *C. elegans* sequence that were shown by our sufficiency analysis to be important for either early expression in the presumptive vulE and vulF cells, or in vulC and vulD cells, respectively (Kirouac and Sternberg, in prep.). Element B resides within a region in *C. elegans* that is important for early expression in the presumptive vulE and vulF cells (Kirouac and Sternberg, in prep.). However, this region alone in *C. elegans* was not sufficient to drive this expression pattern consistently. Element D is in a region in *C. elegans* that was shown by sufficiency analysis to be important for driving vulC and vulD expression (Kirouac and Sternberg, in prep.). Element A and C lie in regions that are not needed to drive vulC and vulD expression in *C. elegans.*

When this analysis was performed at a lower threshold of 85% identity, with *C. elegans* sequence mk80-132 (4316-4474) (Figure 5A) needed to drive expression in vulC and vulD, another region, element E, is identified (Figure 5B).

For the *zmp-1* comparison, we used the *C. elegans* genomic sequence from the region mk50-51 (Figure 6A), which we have shown through sufficiency analysis to be important for vulva expression in vulA, vulE, and the anchor cell, as a basis for comparison against the *C. briggsae* sequence upstream of the predicted *zmp-1* translational start site. This comparison was performed in the same manner as for *egl-17*. At this threshold level, four regions of similarity were found (Figure 4B). The order of these four elements (A, B, C and D) is conserved (Figure 6). However, element D is in the reverse orientation with respect to the other elements and the coding region; element D lies within a region in *C, elegans* that is crucial for anchor cell and vulE cell expression (Kirouac and Sternberg, in prep.). Part of this region was deleted in the Δ3/4 *zmp-1* internal deletion, which shows loss of expression in vulE. The B element is located in a region in *C. elegans* that was shown by deletion analysis to be important for vulA expression (Kirouac and Sternberg, in prep.). Element A appears at the 90% threshold level, while the rest of these elements appear at the 85% level.

For the *cdh-3* comparison, we performed two separate analyses. The first analysis was performed using the upstream region from *C. elegans,* 2290-3419 (mk96-134) (Figure 7A) that was shown to drive both anchor cell expression and vulva cell expression (the first vulval region; Kirouac and Sternberg, in prep.). This sequence was analyzed using the Family Relations and Seqcomp programs to identify regions of similarity when compared to the sequence upstream of the predicted translational start site of *C. briggsae cdh-3*. At a threshold level of 85% identity, six elements where found (Figure 4C). These elements, A-F, are scrambled with respect to each other between the two species, both in location and orientation (Figure 7). Element A resides within the α

region, element B resides within the β region, and element F resides within the γ region defined by the sufficiency analysis in *C. elegans* (Kirouac and Sternberg, in prep.). These three sites are important for anchor cell expression, and may also help drive expression in vulE, F, C, and D (Figure 7). Element F shares 100% identity between the two species, while the rest of these elements share 85% homology. All three of the remaining elements D, E, and F, as well as part of C, are contained in the *C. elegans* region mk118-143 that drives variable expression in the vulD, vulE and occasional vulC cells (Kirouac and Sternberg, in prep.).

The second analysis of *cdh-3* was performed with the *C. elegans* genomic sequence corresponding to the mk66-67 (4434-4997)(Figure 8A), which contains the second region that was sufficient to drive expression in the vulva cells (Figure 4D; Kirouac and Sternberg, in prep.). When this region was compared at an 85% threshold level with the sequence upstream of the predicted translational start of *C. briggsae cdh-3*, four elements were found: H, I, J and K. Again, the order of these elements were scrambled between the two species, and these elements partially overlap (Figure 8). Element K shares 100% identity, elements J and H share 95% identity, and element I shares 85% identity between the two species.

Analysis of *C. briggsae* upstream regions

To assess the role of these conserved elements in the cell-specific regulation of these genes, we made constructs containing the elements found in the upstream region of *egl-17*, *zmp-1* and *cdh-3* in *C. briggsae* (Table 1).

Construct mk160-161 (a 748-bp fragment containing the *C. briggsae egl-17* elements B, C, D and E) (Figure 5B), when injected into *C. elegans*, drives expression in both vulC and vulD cells, as well as early expression in the presumptive vulE and vulF cells (Table 1, and Figure 9A). In all lines examined, animals showed variable early expression. Not only was GFP expressed in the presumptive vulE and vulF cells, but GFP was also expressed in the presumptive vulA, B, C and D cells; this latter expression perdured into later stages of invagination (through L3 in some cases, but never in L4) than in *C. elegans*. Furthermore, GFP was sometimes not expressed in the presumptive vulE and vulF cells, while it was expressed in presumptive A, B, C, and D cells. It is possible that in this construct, a negative regulatory element is missing, thereby giving rise to the expanded expression pattern and extending the duration of expression. It is also possible that this expression pattern is the result of species differences either in regulatory control, or in protein function. Element B, which plays a role in expression in the presumptive vulE and vulF cells, is located ~200 bp upstream of the region that correlates with vulC and vulD expression. However, in *C. elegans* this potential enhancer element is located over 1 kb away from the elements that are driving the vulC and vulD expression (Kirouac and Sternberg, in prep.). This observation suggests that the spacing of these elements may not be critical for their functionality.

The *C. elegans egl-17*::GFP reporter, containing 3.9 kb of upstream sequence, shows the same expression pattern in *C. briggsae* as it does in *C. elegans* (Table 1). An occasional animal does not express GFP in vulC and vulD cells at the L4 stage. However, when the 748 bp construct mk160-161 was injected into *C. briggsae*, expression was not seen in the presumptive vulE and vulF cells at the VPC 4-cell stage, although an

occasional animal that was starting to invaginate did show expression in P5.p (Table 1). This observation suggests that either all the elements required for the fidelity of the early expression in *C. briggsae* are not contained in this construct, or that the native gene in *C. briggsae* is not expressed in these cells. In L4 animals, GFP was expressed in the vulva in about 50% of the animals. Of this 50%, GFP was consistently expressed in vulC, and sometimes vulD cells. We infer that an element that is necessary for the fidelity of the expression in *C. briggsae* in vulC and vulD may be missing. Furthermore, this missing element plays a proportionally larger role in regulating the expression in vulD than in vulC cells.

Construct mk172-173 (5138-5892), a 755 bp fragment containing the *C. briggsae zmp-1* elements A, B, C and D (Table 1 and Figure 6B), when injected into *C. elegans*, drives expression in the anchor cell, vulE and vulA (data not shown). The only apparent difference between the expression pattern in *C. elegans* and *C. briggsae* is that the vulA expression is variable, and seems to occur at slightly later time points. This difference suggests that there may be an additional element(s) not present in mk172-173 that ensures the fidelity of the vulA expression. In *C. elegans*, vulA expression can be seen in the young adult, but mk172-173 drives vulA expression slightly later than its *C. elegans* counterpart; the majority of animals do not express GFP in vulA cells until eggs are present in the uterus.

The *C. elegans zmp-1*::GFP reporter, containing 3.5 kb of upstream sequence, shows the same expression pattern in *C. briggsae* as it does in *C. elegans* (Table 1). Consistent expression was seen in the anchor cell, vulA and vulE. Expression in vulD cells in *C. briggsae* was not determined because of its weak expression in *C. elegans*.

A 1.4-kb fragment containing the *C. briggsae cdh-3* elements A, B, D, E and F (mk162-163) (Figure 7B), when injected into *C. elegans*, drives expression in the vulE, F, C and D cells, but less than 10% of the animals showed any expression in the anchor cell (Table 1 and Figure 9B). A similar fragment, mk96-134 (2290-3419) (Figure 7A) from *C. elegans*, drives expression in vulE, F, C, D and anchor cell (Kirouac and Sternberg, in prep.).

A 277 bp fragment containing the *C. briggsae cdh-3* elements H, I, J, and K (mk164-165) (Table 1 and Figure 8B), when injected into *C. elegans*, drives expression in vulC, E, and F, but not in vulD (data not shown). This expression pattern varies from animal to animal, with vulF showing the strongest and the most penetrant expression. A similar fragment, mk66-67 (4434-4997) (Figure 8A), from *C. elegans*, drives expression in vulE, F, C and D cells (Kirouac and Sternberg, in prep.).

The *C. elegans cdh-3*::GFP reporter, containing 6.0 kb of upstream sequence, does not show the same expression pattern in *C. briggsae* as it does in *C. elegans* (Table 1). Although the expression in the anchor cell is present consistently, only rarely is there expression in vulC, D, E, or F. When there is expression in the vulva cells, it is usually not present in more than a single cell in any given animal. This is in spite of the fact that when the *cdh-3 C. briggsae* sequences are placed in the context of *C. elegans*, there is some expression in vulval cells. We infer that the factor(s) that drive expression in *C. elegans* might be absent in the corresponding *C. briggsae* cells, or the factors have altered binding specificity in *C. briggsae*. It is possible that this gene may have different functions in these two species. Alternatively, *C. briggsae cdh-3* may use binding sites not present in the 6.0 kb of the *C. elegans* sequence to drive expression in the vulva cells.

Transfac binding site prediction in conserved regions

As one approach to finding potential binding sites for known transcription factors in the conserved region, we used the MatInspector program (http://www.genomatix.de/mat_fam; Quandt *et al*., 1995). We set the core matrix similarity to a minimum of 0.90 to maximize the specificity of the binding sites. We then compared the output from the program of *C. elegans* mk84-148 (3182-4732) to the output for the *C. briggsae* mk160-161 (17543-18289). Only binding sites that appear in both these sequences and had a maximum Random Expectation Value (re-value; the "re" value is the number of times the sequence would appear by chance in 1000 bp of sequence) of ≤ 0.51 were considered for further analysis (Table 2). This process was repeated to compare *C. elegans* sequences from mk96-134 (2290-3419) to *C. briggsae* sequences for mk162-163 (22710-21306), and *C. elegans* sequences for mk66-67 (4434-4962) to sequences for *C. briggsae* construct mk164-165 (18143-17867). Finally, this analysis was done for *C. elegans* construct mk50-51 (1052-1438), and to sequences for *C. briggsae* construct mk172-173 (5138-5892). A total of four potential binding sites were found in the conserved regions of *egl-17* (Table 2). All four of these sites were located in element D. *zmp-1* contained eight factor binding sites in conserved regions (all located in conserved region B or D). The first *cdh-3* region containing conserved elements A-F had three factor binding sites in conserved regions (located in elements B and F; Table 2), and the second *cdh-3* region containing elements H-K also had three conserved binding sites (all located in element K; Table 2). Although this program predicted putative binding sites for families thought to play a role in the specification or terminal differentiation of

these cells (e.g. ETS family members, TCF/LEF-1), we found only two putative binding sites for factors from these families whose site is located in one of the conserved regions of *C. elegans*, and whose corresponding element in *C. briggsae* also contains the same site. The first family was the LIM homeodomain family; *lin-11* is a LIM domain family member and is known to play a role in the specification of secondary cells (Freyd *et al*., 1990). LIM domain family member sites are found in conserved regions of *egl-17* and *cdh-3* (mk66-67/ mk164-165 region). The second family is the HOX homeodomain family (Kenyon *et al*., 1998). There is a conserved site in *cdh-3* (mk96-134/ mk162-163) and *zmp-1* (mk50-51/ mk172-173). However, the consensus for the homeodomain families is very weak outside the TAAT core. Given the low specificity, we did not mutate these sites.

AlignACE predictions of overrepresented sequences

We used the AlignACE program (Roth *et al*, 1998), which computes motifs based on sequences that are over-represented in the input sequence, to identify motifs in the upstream sequences of the *C. briggsae egl-17*, *zmp-1* and *cdh-3* (Table 3). We then looked to see which of those motifs were localized in conserved elements. We chose this approach instead of searching for common motifs between homologous upstream regions, because homologous upstream regions, by definition, are likely to be more similar. While looking for regions of similarity was an effective approach to identifying important regulatory sequences within a large upstream sequence, the Seqcomp and Family Relations programs (Brown *et al*., 2002) recognizes matches based on 85%-100% percent identity over a window of 20 base pairs. The AlignACE program identifies motifs

based on a consensus of eight to ten base pairs. These matches will likely occur much more frequently between two homologous upstream regions than those in two coregulated genes, and may not be functionally meaningful. We also searched for motifs that were common to *C. briggsae zmp-1* region mk172-173 and *C. elegans cdh-3* region mk96-134, each of which are sufficient to drive expression of a naïve promoter in the anchor cell.

In our analysis of *C. briggsae egl-17* sufficiency region mk160-161, AlignACE identified three 8-bp motifs and two 10-bp motifs above the threshold MAP cut-off of 10 (Table 3A). Several of these motifs have common sites, which suggests that they are either variants of the same motif or that they might represent binding sites of trans-acting factors that cooperatively bind DNA. Motifs 1.8, 2.8, 3.8 and 5.10 all have roughly the same site in conserved element B, which was implicated in a sufficiency analysis to be important for the fidelity of the early expression in the presumptive vulE and vulF cells (Table 3B; Kirouac and Sternberg, in prep.). In addition, all of the motifs except 5.10 had sites that resided within conserved element D; element D is located in a region that is critical for conferring expression in vulC and vulD cells (Table 3B; Kirouac and Sternberg, in prep.).

The analysis of the *C. briggsae zmp-1*region mk172-173 identified three 8-bp motifs and two 10-bp motifs (Table 3A). While motifs 1.8, 3.8, and 4.10 all contained sites in conserved element D, only motif 1.8 was found within the part of this element that is contained in the sufficiency region mk50-51 in *C. elegans* (Table 3B; Kirouac and Sternberg, in prep.). It is possible that conserved element D plays a role in conferring expression in vulA cells. Motif 5.10 has one site that is found in conserved element A;

conserved element A is a region that was shown by sufficiency analysis to be critical for anchor cell expression in *C. elegans* (Table 3B; Kirouac and Sternberg, in prep.).

In *C. briggsae cdh-3* construct mk162-163 nine 8-bp motifs and five 10-bp motifs were identified (Table 3A). Of these motifs, 4.8, 5.8, 7.8, 8.8, 10.10, and 12.10 each had one site in conserved element F. This element is in a region that by sufficiency analysis in *C. elegans* was important for both vulval and anchor cell expression (gamma region, Kirouac and Sternberg, in prep.). Motifs 8.8, 12.10 and 13.10 all contain a site in conserved element D, and a site in conserved element A (Table 3B). It is unclear at this time what role conserved element D might be playing in regulating *cdh-3* expression. Conserved element A is located in the alpha region that is important for anchor cell expression in *C. elegans*, but mk162-163 was not able to drive expression in the anchor cell except in few rare cases. Element A's role, if any, in driving expression in vulval cells is not evident.

Mk164-165 was examined using the AlinACE program and was found to contain one 8-bp and two 10-bp motifs (Table 3A). Taken together, these motifs have sites in conserved elements H, J K and I. Mk164-165 drives vulE, F, C, but not D cell expression in *C. elegans* vulval cells (Table 3B, Kirouac and Sternberg, in prep.). The conservation through this region is extensive, suggesting that these regions of conservation and, as an extension of this, these motifs may be important in conferring this expression.

We also compared the *C. briggsae zmp-1* mk172-173 to the *C. elegans cdh-3* mk96-134; both of these regions are sufficient to confer anchor cell expression on a naïve promoter. AlignACE was able to identify one 8-bp motif and two 10-bp motifs that scored above the MAP score cut-off of 10 (Table 3A). An ideal candidate motif would

have sites in conserved regions of both *cdh-3* and *zmp-1* (in essence giving a four-way comparison). Unfortunately in this case, while all three motifs have at least one site that is located in conserved element A of the *cdh-3* region, no sites fall in the conserved elements identified in *zmp-1* (Table 3B). We did not do the reciprocal comparison since the *C. briggsae* construct, which contains the conserved elements that appear to be important in conferring anchor cell specificity in *C. elegans*, does not drive expression in the anchor cell in *C. elegans*.

DISCUSSION

Experiments testing the sufficiency of genomic fragments to direct expression of a heterologous promoter defined small regions that are critical for the fidelity of the expression pattern of *C. elegans egl-17, zmp-1* and *cdh-3* (Kirouac and Sternberg, in prep.). However, these regions were still too large to identify specific putative binding sites for known transcription factors. In order to further experimentally define possible binding sites for transcription factors, we used phylogenetic footprinting of the cis-regulatory regions between two species of *Caenorhabditis*, *C. elegans* and *C. briggsae*: *C. briggsae*, by molecular criteria, is 50-120 million years diverged from *C. elegans* (Coghlan, 2002). The Seqcomp program (Brown *et al*., 2002) was crucial in identifying conserved elements between *C. elegans* and *C. briggsae* in upstream regions. By using phylogenetic footprinting in homologous genes in addition to correlating putative binding sites in potentially co-regulated genes (Kirouac and Sternberg, in prep.), we have maximized the likelihood of identifying regulatory elements responsible for cell-type specific expression.

Phylogenetic footprinting

When phylogenetic footprinting is carried out on a whole-genome scale, it identifies the most highly conserved elements in the regulatory regions; these are promising candidates for binding trans-acting factors (reviewed in Blanchette and Tompa, 2002). In our analysis, we already had in our hands relatively small regions from the homologous *C. elegans* genes that were sufficient to direct vulva and anchor cell expression (Kirouac and Sternberg, in prep.). In the case of *egl-17*, there was a coincidence of the conserved region with the functionally defined sequences at the 95-90% identity level; there were only four elements that were conserved in the 3.9 kb of the original reporter construct. However, for both *cdh-3* and *zmp-1*, there were many conserved elements that did not necessarily fall in the realm of the previously defined sufficiency pieces (Kirouac and Sternberg, in prep.). In *zmp-1*, at a threshold level of 85% identity, there are two to four blocks of conservation in the upstream regions. One of these blocks is the region around mk50-51. In *cdh-3* at a threshold level of 100% identity, three conserved regions appear; elements K and F are two of these three regions. At a threshold level of 90%, element K expands as does the third site, and one additional region appears. Finally, at the 85% threshold level, we see multiple sites spread out throughout the upstream region. This fact made the sufficiency data invaluable for determining which of these conserved elements may play a role in directing vulva and anchor cell specificity. It seems likely that these other conserved regions may be conserved elements involved in the regulation of this gene in other tissues. *egl-17* ::GFP is expressed in a limited number of other tissues: in two large unidentified cells in the head at the three-fold stage of embryogenesis, in the

M4 pharyngeal neuron, and occasionally in the ventral hypodermis of late first-stage larvae (Burdine *et al*., 1998). In *C. elegans*, *zmp-1*:GFP is expressed in a variety of other cell types, from multiple lineages, including uterine and tail cells, and body muscle and subsets of neurons (J. Butler and J. Kramer, unpublished data). In hermaphrodites, *cdh-3*::GFP is expressed in the seam cells, the buccal and rectal epithelia, the excretory cell, two hypodermal cells in the tail, the uterine epithelium closest to the invaginating vulval cells followed by the multinucleate uterine seam cell (utse), the vulva and associated neurons (Pettitt *et al*., 1996). The complexity of the expression patterns, and the variety of tissues in which both *zmp-1* and *cdh-3* expression are expressed contrasts with the relatively simple expression pattern of *egl-17*::GFP, thus these other conserved regions in *zmp-1* and *cdh-3* may be other cis-regulatory regions driving transcription in other tissues. It may also be the case that some genes have undergone a faster rate of divergence than others have, and may be under less selective pressure.

Potential for specific isolation of trans-acting factors binding sites by phylogenetic footprinting between *C. elegans* and *C. briggsae*

By comparing the phylogenetic footprints in the upstream regions of homologous sequences from *C. elegans* and *C. briggsae*, we were able to narrow down regions that were responsible for the vulva and anchor cell specific expression of these genes. However, we could not determine distinct binding sites. Cis-regulatory binding sites can be eight to 10 bp long and they are often highly variable; since DNA has only four-fold variation instead of the 20-fold seen in protein, its level of random variation can be quite high. Comparison to *C. briggsae* will be helpful in locating a phylogenetic footprint of

conserved regulatory regions and confirming the presence of a putative binding site(s). However, when there are no obvious trans-acting candidates, it may be necessary to compare co-regulated or homologous genes from several other species to detect signal above background.

Analysis of putative trans-acting factors using the Transfac database

The focus of these studies was to isolate cell-specific cis-regulatory response elements. However, we also used the Transfac database to look for putative trans-acting factors in the conserved regions that drive expression in the anchor and vulva cells (Table 2), and to compare these data to the putative binding sites in found in the sufficiency analyses (Kirouac and Sternberg, in prep.). Putative binding sites in the conserved elements between *C. elegans* and *C. briggsae* upstream sequences overlap with only a few putative sites defined by the sufficiency analysis of these potentially co-regulated genes (Kirouac and Sternberg, in prep.). Among the overlap were: the CLOX family members, CDP and CDPCR3; the glucocorticoid response family member, GRE; the octamer family member, Oct1; and the homeodomain proteins ISLI and MEIS-1. It is likely that the expression is driven in these cells by different combinations of factors, and that we will not be able to isolate a factor(s) responsible for driving the expression in a single cell type across a panel of coregulated genes, or in orthologous genes in different species.

While a number of genes (for example, *egl-38*, *lin-26*, *lin-29*, cog-*1* and *lin-11*) (Freyd *et al.,* 1990; Labouesse *et al.*, 1994; Rougvie and Ambros, 1995; Bettinger *et al.*, 1997; Chamberlin *et al.*, 1997; Palmer *et al*., in press) are known to effect the marker gene expression patterns in the vulva, it is not yet known whether they act directly in the

regulation of these genes, or more proximally in the specification of these cell types (M. Wang, T. Inoue, and P. Sternberg, unpublished data). Of these genes, only a site potentially bound by *lin-11* showed up in our Transfac analysis. Biochemical studies using the sufficiency pieces and the conserved regions defined in these studies might help determine which of these transcription factors has a direct effect on the transcriptional regulation of these genes.

Analysis of over-represented sequences in regions of sufficiency

While the Transfac database (Quandt *et al*., 1995) identifies binding sites of known transcription factors, AlignACE (Roth *et al*, 1998) identifies sequences that are over-represented in a given sequence. This approach allows the isolation of candidate motifs either within a gene, or between genes. We were able to use this program to identify motifs in our *C. briggsae* constructs, and evaluate whether these motif sites resided in any of the conserved regions that were found using the Seqcomp and Family Relations programs. When we compared *C. briggsae* mk172-173 and *C. elegans* 96-134, each of which are expressed in the anchor cell, we were able to isolate several motifs that may be binding sites of factors that play a role in conferring this cell-specific expression.

Implications of cross-species comparison of *egl-17*, *zmp-1* and *cdh-3*

By comparing the expression patterns of the full-length *C. elegans* GFP reporter constructs in *C. elegans* and *C. briggsae*, it appears that there might be inter-species differences in gene regulation and function. Both *egl-17* and *cdh-3* show differences in expression patterns in the vulva and anchor cell in *C. briggsae*.

The *C. elegans egl-17*::GFP reporter, containing 3.9 kb of upstream sequence, shows expression in the same vulval cells in *C. briggsae* as it does in *C. elegans*. However, there are some differences. Occasionally, *C. briggsae* animals do not express *egl-17*::GFP in vulC and vulD cells at the L4 stage. It is unknown whether this is a result of DNA-mediated transformation differences between *C. elegans* and *C. briggsae*, or if it reflects differences in gene regulation. Early expression is grossly the same between the two species when we examined the full-length *C. elegans* construct in *C. briggsae*. However, when the *C. briggsae egl-17* conserved upstream sequence mk160-161 was injected into *C. elegans*, early expression was highly variable, and was driven in P5.p and P7.p and their descendants as often as it was driven in P6.p. This same region, when injected into *C. briggsae*, does not show consistent expression in the primary lineage, but does show occasional expression in the secondary lineage, P5.p. This difference suggests that there may be a repressor site in *C. elegans* that inhibits expression in vulval cells outside of the primary lineage. However, occasionally, in *C. elegans*, the *C. elegans egl-17*::GFP expression is observed in the secondary lineages at the VPC four-cell stage, but this expression is always in addition to expression in P6.p (M. Wang, D. Sherwood and M. Kirouac, unpublished observations).

While, the differences in the *egl-17*::GFP expression pattern may only be the result of quantitative differences in binding specificity of one or more transcription factors, the differences in *cdh-3*::GFP expression are more substantial. These differences indicate that *cdh-3* may be playing a different role in the vulval cells in *C. briggsae*. In *C. elegans* it is clear that CDH-3 is required for the morphogenesis of a single cell that forms the tip of the tail in the hermaphrodite, while the other cells that express the *cdh-3*

reporter appear to be unaffected by a null allele (Pettitt *et al*., 1996). However, the genesis of the egg-laying system requires several sets of cell-cell recognition events, all of which occur during the expression of *cdh-3*::GFP. The vulval epidermal cells invaginate and form a connection with the uterus, and the utse cell makes contacts with the seam cells. In addition, during the formation of the seven toroidal rings of the vulva, the vulva cells are involved in complex interactions (Pettitt *et al*., 1996; Sharma-Kishore *et al*., 1999). It is possible that in *C. elegans,* other genes can compensate for the loss of CDH-3. There are 12 predicted cadherin superfamily members in *C. elegans*. Of these 12, two, *hmr-1* and *cdh-3,* have been defined by experimental work on their structure and function (Tepass, 1999). Since it appears that in *C. elegans*, *cdh-3* is not required in the vulva cells, it is even less clear what is going on in *C. briggsae.* Perhaps, in *C. briggsae* other members of the cadherin family are active in the vulva cells, or perhaps this gene family is not active at all in the *C. briggsae* vulva.

Conclusions

Independent analysis by phylogenetic footprinting and sufficiency testing (Kirouac and Sternberg, in prep.) can define similar control regions for conferring cell-type specific expression (e.g., regions that drive *egl-17* expression in the vulval cells can be found independently by both methods). However, the success of *de novo* analysis using phylogenetic footprinting techniques will likely depend on the complexity of the cis-regulatory control region. The more complex the control region, the more one must rely on other data, such as sufficiency testing, in establishing the appropriate region for any given cell-type specific expression. In our study, both the *zmp-1* and *cdh-3* upstream

regions had multiple regions of similarity, and it was only through the use of our sufficiency data that we were able to correctly identify regions that conferred vulval cell and anchor cell expression. While these modules may not be narrow enough to resolve discrete binding sites, the addition of other species may allow sub-domains of these phylogenetic footprints to be identified and tested for their ability to confer cell-type specific expression. Also, we found evidence of differences in the expression of both *egl-17* and *cdh-3* full-length *C. elegans* reporter constructs in *C. briggsae*; such differences suggest that either the regulation, or the function, or both, of these proteins has changed in the last 50-120 million years. The convergence of cross-species sufficiency studies and phylogenetic footprinting studies is an efficient way to identify candidate factor binding sites.

Acknowledgments

We are grateful to Jim Butler and Jim Kramer for *zmp-1*::GFP and the sharing of unpublished data; Rebecca Burdine and Micheal Stern for *egl-17*::GFP; and Jonathan Pettitt, William B. Wood and Ronald Plasterk for the *cdh-3*::GFP. We would like to thank Cheryl Van Buskirk, Takao Inoue, Erich Schwarz, Gary Schindelman, and Bhagwati Gupta for the critical reading of this manuscript. P.W.S is an investigator with the Howard Hughes Medical Institute, which supported this research.

REFERENCES

Bettinger, JC., Euling, S., and Rougvie, AE. (1997). The terminal differentiation factor LIN-29 is required for proper vulval morphogenesis and egg laying in *Caenorhabditis elegans*. *Development* **124**, 4333-4342.

Blanchette, M., and Tompa, M. (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research* **12,** 739-748.

Brown, C., Rust, A., Clarke, P., Pan, Z., Schilstra, M., De Buysscher, T., Griffin, G., Wold, B., Cameron, R., Davidson, E., and Bolouri, H. (2002). New Computational Approaches for Analysis of cis-Regulatory Networks. *Developmental Biology* **246,** 86-102.

Burdine, R., Branda, C., and Stern, M. (1998). EGL-17(FGF) expression coordinates the attraction of the migrating sex myoblasts with vulval induction in *C. elegans*. *Development* **125,** 1083-1093.

Chamberlin, HM., Palmer, RE., Newman, AP., Sternberg, PW., Ballie, DL. and Thomas, JH. (1997). The PAX gene *egl-38* mediates developmental patterning in *Caenorhabditis elegans*. *Development* **124,** 3919-3928.

Coghlan, A. W. K. (2002). Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. *Genome Research* **12,** 857-867.

Costanzo, M., Hogan, J., Cusick, M., Davis, B., Fancher, A., Hodges, P., Kondu, P., Lengieza, C., Lew-Smith, J., Lingner, C., Roberg-Perez, K., Tillberg, M., Brooks, J., and Garrels, J. (2000). The Yeast Proteome Database (YPD) and *Caenorhabditis elegans* Proteome Database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Research* **28,** 73-76.

Culetto, E., Combes, D., Fedon, Y., Roig, A., Toutant, J., and Arpagaus, M. (1999). Structure and promoter activity of the 5' flanking region of *ace-1*, the gene encoding acetylcholinesterase of class A in *Caenorhabditis elegans*. *Journal of Molecular Biology* **290,** 951-966.

de Bono, M., and Hodgkin, J. (1996). Evolution of sex determination in *Caenorhabditis*: Unusually high divergence of *tra-1* and its functional consequences. *Genetics* **144,** 587-595.

Fodor, A., Riddle, D., Nelson, F., and Golden, J. (1983). Comparison of a new wild-type *Caenorhabditis briggsae* with laboratory strains of *C. briggsae and C. elegans*. *Nematologica* **29,** 203-217.

Freyd, G., Kim, S., and Horvitz, H. (1990). Novel cysteine-rich motif and homeodomain in the product of the *Caenorhabditis elegans* cell lineage gene *lin-11*. *Nature* **344,** 876-879.

Gilleard, J., Henderson, D., and Ulla, N. (1997). Conservation of the *Caenorhabditis elegans* cuticle collagen gene *col-12* in *Caenorhabditis briggsae*. *Gene* **193,** 181-186.

Gower, N., Temple, G., Schein, J., Marra, M., Walker, D., and Baylis, H. (2001). Dissection of the promoter region of the inositol 1,4,5-triphospate receptor gene, *itr-1*, in *C. elegans*: A molecular basis for cell-specific expression of IP3R isoforms. *Journal of Molecular Biology* **306,** 145-157.

Greenwald, I. (1997). Development of the Vulva in: *"C. elegans II." DL Riddle, T Blumenthal, BJ Meyer and JR Priess (eds), Cold Spring Harbor Laboratory Press.* **II,** 519-541.

Higgins, D.G., Thompson, J.D., and Gibson, T.J. (1988). Using ClustalW for multiple alignments. *Methods Enzymol.* **266**, 387-402.

Hughes, JD., Estep, PW., Tavazoie, S., and Church, GM. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* **296**, 1205-1214.

Kennedy, B., Aamodt, E., Allen, F., Chung, M., Heschl, M., and McGhee, J. (1993). The gut esterase gene (*ges-1*) from the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Journal of Molecular Biology* **229,** 890-908.

Kenyon, C., Austin, J., Costa, M., Cowing, D., Harris, J., Honigberg, L., Hunter, C., Maloof, J., Muller-Immergluck, M., Salser, S., Waring, D., Wang, B., and Wrischnik, L. (1998). The dance of the Hox genes - Patterning the anteroposterior

body axis of *Caenorhabditis elegans*. *Cold Spring Harbor Symposia on Quantitative Biology* **62,** 293-305.

Kornfeld, K. (1997). Vulval development in *Caenorhabditis elegans*. *Trends in Genetics* **13,** 55-61.

Krause, M., Harrison, S., Xu, S., Chen, L., and Fire, A. (1994). Elements regulating cell- and stage-specific expression of the *C. elegans* myoD family homolog *hlh-1*. *Developmental Biology* **166,** 133-148.

Kuwabara, P. (1996). Interspecies comparison reveals evolution of control regions in the nematode sex-determining gene *tra-2*. *Genetics* **144,** 597-607.

Labouesse, M., Sookhare, S., and Horvitz, HR. (1994). The *Caenorhabditis elegans* gene *lin-26* is required to specify the fates of hypodermal cells and encodes a presumptive zinc-finger transcription factor. *Development* **120**, 2359-2368.

Maduro, M., and Pilgrim, D. (1996). Conservation of function and expression of unc-119 from two *Caenorhabditis* species despite divergence of non-coding DNA. *Gene* **183,** 77-85.

Mello, C., Kramer, J., Stinchcomb, D., and Ambros, V. (1991). Efficient gene transfer in *C. elegans*: Extrachromosomal maintenance and integration of transforming sequences. *EMBO Journal* **10,** 3959-3970.

Nigon, V., and Dougherty, E. (1949). Reproduct patterns and attempts at reciprocal crossing of *Rhabditis elegans* Maupas, 1900, and *Rhabditis briggsae* Dougherty & Nigon, 1949 (Nematoda: Rhabditidae). *Journal of Experimental Zoology* **112,** 485-503.

Pettitt, J., Wood, W., and Plasterk, R. (1996). *cdh-3*, a gene encoding a member of the cadherin superfamily, functions in epithelial cell morphogenesis in *Caenorhabditis elegans*. *Development* **122,** 4149-4157.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* **23,** 4878-4884.

Roth, FP., Hughes, JD., Estep, PW., and Church, GM. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology* **10**, 939-945.

Rougvie, AE., and Ambros, V. (1995). The heterochronic gene *lin-29* encodes a zinc finger protein that controls a terminal differentiation event in *Caenorhabditis elegans*. *Development* **121**, 2491-2500.

Sharma-Kishore, R., White, J., Southgate, E., and Podbilewicz, B. (1999). Formation of the vulva in *Caenorhabditis elegans*: a paradigm for organogenesis. *Development* **126,** 691-699.

Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J. (2001). WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Research* **29,** 82-86.

Stern, D. L. (2000). Evolutionary developmental biology and the problem of variation. *Evolution Int J Org Evolution* **54,** 1079-91.

Sternberg, P., and Han, M. (1998). Genetics of RAS signaling in *C. elegans*. *Trends in Genetics* **14,** 466-472.

Tagle, D., Koop, B., Goodman, M., Slightom, J., Hess, D., and Jones, R. (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal of Molecular Biology* **3,** 439-455.

Tepass, U. (1999). Genetic analysis of cadherin function in animal morphogenesis. *Current Opinion in Cell Biology* **11,** 540-548.

Wang, M., and Sternberg, P. (2000). Patterning of the *C. elegans* primary vulval lineage by RAS and Wnt pathways. *Development* **127,** 5047-5058.

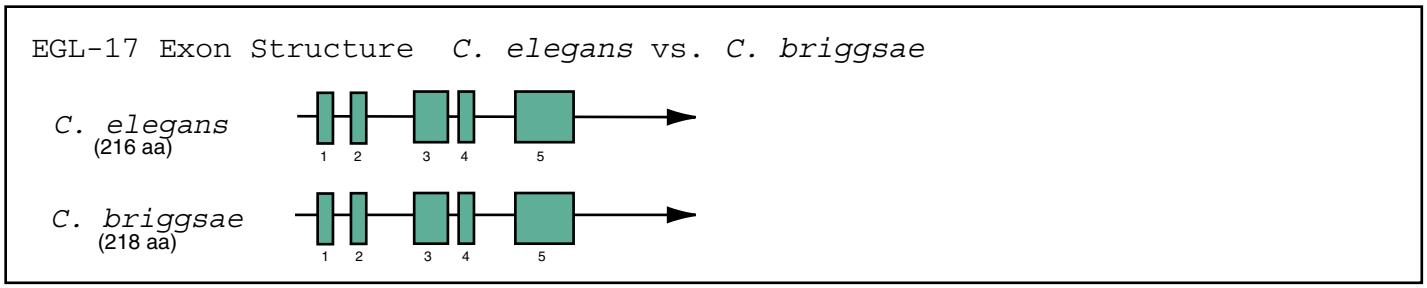
Webb, C., Shabalina, S., Ogurtsov, A., and Kondrashov, A. (2002). Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Nucl. Acids Res* **30,** 1233-1239.

Xue, D., Finney, M., Ruvkun, G., and Chalfie, M. (1992). Regulation of the *mec-3* gene by the *C. elegans* homeoproteins UNC-86 and MEC-3. *EMBO Journal* **11,** 4969-4979.

Figure 1: EGL-17 clustalW alignment in *C. elegans* and *C. briggsae*

The exon structures are shown at the top of the figure. The *C. elegans egl-17* has one untranslated exon that is not shown in the exon structure. The exon that starts with the translational start is labeled exon 1. Exon boundaries are indicated by an inverted triangle. The *C. briggsae* cDNA corresponding to the amino acids highlighted in blue was sequenced from a RT-PCR. In this alignment, * indicates amino acid identity, : identifies a highly conserved amino acid substitution, and . indicates there is a semi-conserved amino acid substitution. The red boxes show the location of the six beta strands, and the green boxes show the location of the three hairpin regions that together make up the beta-trefoil fold, which is conserved in the FGF ligand family.

Figure 1: EGL-17 clustalW alignment *C. elegans* and *C. briggsae*

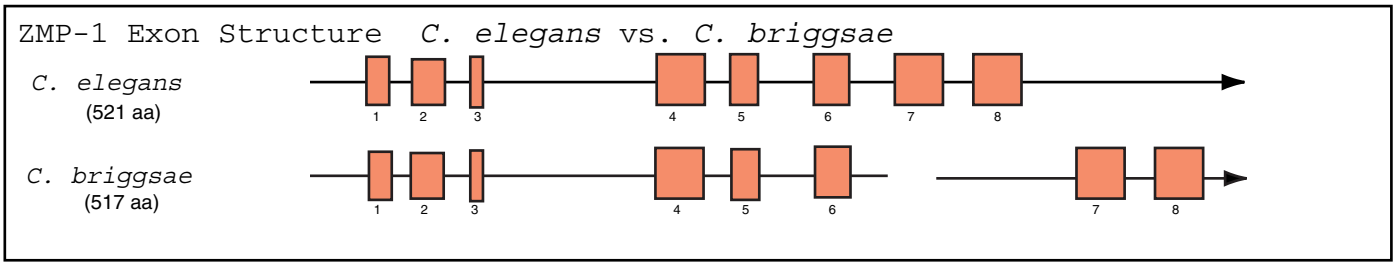



```
C. elegans      MLKVLLTLMLSTNFRNTCARFQIKPNVHYLGETHWQLFNECSQGMLQSFLGSLNTRGYPD 60
C. briggsae     MLDILFILLMS-NAGHTCARFNMKANVHHIGETHWQLFNECSKGMLQSFLGSLNTRGYPD 59
                **.:*: *::*:*  :*****::*.***::**********:***************

C. elegans      KHCLTDWNVVGEWDGKFRLQHAQSRKFLCFNKRARITLRFNGSDAKCTFIEEVRDNGFSR 120
C. briggsae     RHCLTDWNVLGEWDGKFRIQHAQSKKFLCFNKRARVTLRFNGSDVKCTFIEEIHENGYSR 119
                :********:*********:******:**********:********.*******:::**:**

C. elegans      LRSSWKPELYLGFNGRGRFQNPLSYHLKPRCFDWIKLVRYVAESEKSVCSTPPKPKLSPS 180
C. briggsae     LRSSWKPELYLGFNSRGRFQNPLSFHLKPRCFDWIKLVRYVPESEKNVCSAPPKPKPS-T 178
                **************.*********:****************.****.**:***** *.:

C. elegans      PLEHSSFVHHAVRSNFLKKVSATHDSLYRTMKSRKS 216
C. briggsae     PIEHSPFAYKVARSHFLKKVSATHESLYR-FTSLKI 213
                *:***.*.::..**:*********:****::.* *
```

Figure 2: ZMP-1 clustalW alignment in *C. elegans* and *C. briggsae*

The exon structures are shown at the top of the figure. The exon begins with the translational start is labeled exon 1. Exon boundaries are indicated by an inverted triangle. The *C. briggsae* cDNA corresponding to the amino acids highlighted in purple was sequenced from a RT-PCR. In this alignment, * indicates amino acid identity, : indicates a highly conserved amino acid substitution, and . indicates a semi-conserved amino acid substitution. The location of the conserved PRCGXPD and HEXXH domains of the matrix metalloproteinase family is shown in black boxes.

Figure 2: ZMP-1 clustalW alignment in *C. elegans* and *C. briggsae*

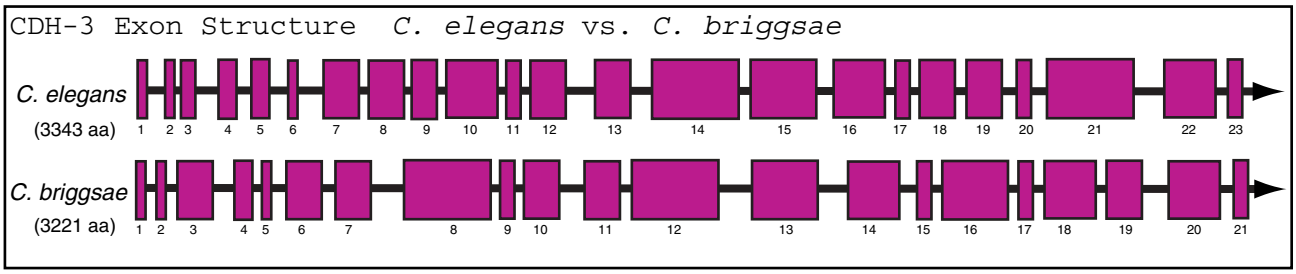



```
C. elegans  MFTGLHDILIILFLLVTLKIAQNVDHTKFLQKYGYLTSGDNQLSSESLSDALKNMQRMAG 60
C. briggsae MFH--YDLLVVFLTIIPFLDTRNVDHTEFLQKYGYLPRGSNQLSSTSLSEALKNMQRMAG 58
            **   :*:*:::: ::.:  ::*****:********. *.***** ***:**********

C. elegans  LEETGELDERTIQMMERPRCGHPDVEDHQKSRGKRYAPPQFKWKEKIITYGCKAVGTSTR 120
C. briggsae LEETGELDERTKRMMERPRCGHPDVQEDRHSRGKRYAQPQFKWTDKVITYGCKSAGTSTR 118
            ********** :*********** ::.::******* ***** .:*:******:.*****

C. elegans  ISLDDLRRTMHQAASQWSELADVEIVESSVKNPMMVISAGRENHYPCTVRFDTKTLAHAF 180
C. briggsae ISLDDLRRTMHQAASQWSELADVEILESSVEKPNIQISAGRSSHYPCNVRFDSQTLAHAF 178
            **************************:****::* : ******..****.****::******

C. elegans  FPTNGQIHINDRVQFAMTNYTERMGANSLYSVVAHEMGHALGFSHSPDIDSVMFAYDTPR 240
C. briggsae FPPHGQIHINDNVHFVMTNFTERIGGNSLYSVVAHEMGHALGFSHSPDSDSVMFAYDTPR 238
            **.:*******.*:*.***:***:*.********** ********** ***********

C. elegans  KWKFTSMDKYNMRSYYGAKASKKENEEEERKTENEDKRRKTEKDRGRTREHESDDIRPNE 300
C. briggsae EWKFTPMDKYRMEMYYGTKKNPTKPKEDERKKE---ERKKEERGRKTEKEHEKDDIRPNE 295
            :****.****.*. ***:* . .: :*:***.*   :*:* *:.*   :***.*******

C. elegans  CRVENPIVVQYRGEYLIFKSQWVWRVSSDWKRLIIKAVPINQLFPGLPNPIDAAVTVGHN 360
C. briggsae CRVDNPIVVQYRGEYLVFKSQWVWRVSSDWKRLIIKPVPIDQLFPGLPNPIDAAVSVGHH 355
            ***:************:*******************.***:**************:***:

C. elegans  LWVFVGEMIYVIYGNHMVHAPLRLSDIGINEKYVDLAYEWHYFNPPAVYIWKGSRYWKLD 420
C. briggsae LWVFVGEKIYVIYGNRMAHAPLTLSEIGIDEKYIDLAYEWHYFNPPAIYIWKGSRYWKLD 415
            ******* *******:*.**** **:***:***:*************:************

C. elegans  EKMYHRRVDERYPKDTDLNWARVPKGVHSAFTYEKEIHLLRGNQVFRMNSSRSVFDIADG 480
C. briggsae EKMDHRRVDERYPKAIDLNWARVPKGVHSAFTYEKEIHFTRGDQVFRMNSSRSVFDVADG 475
            *** **********  **********************: **:**************:***

C. elegans  YPQPLQSFFGFCPRNEKLVLNSSSSHFSLIYATITILILIF-- 521
C. briggsae YPQPLQSFFGFCPRNEKLVLN-ISPTFHPITALLTVLILNFYL 517
            ********************  *. *  * * :*:*** *
```

Figure 3: CDH-3 clustalW alignment in *C. elegans* and *C. briggsae*

The exon structures are shown at the top of the figure. The exon begins with the translational start is labeled exon 1. The lower panel shows the alignment of the first few exons of CDH-3. Exon boundaries are indicated by an inverted triangle, and an inverted triangle with an apostrophe means that an exon boundary was found only in the *C. elegans* protein. The *C. briggsae* cDNA corresponding to the amino acids highlighted in green was sequenced from a RT-PCR. In this alignment, * indicates amino acid identity, : indicates a highly conserved amino acid substitution, and . indicates a semi-conserved amino acid substitution. The conserved cadherin domains of the cadherin family located in this part of CDH-3 are located in the black boxes.

Figure 3: CDH-3 clustalW alignment in *C. elegans* and *C. briggsae*



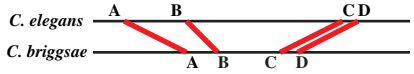

```
C. elegans      MTIRIFFSIFLLNHLIFFHLFNFTHQFSEETIKFSVSEDAKLNTIIGHLEAEIGYTYRLS 60
C. briggsae     ----MLIRHHLFLVFLLTLIFKFSRQFSEETVKFSIAEDAPIDTIIGHLTPENGYSYRLS 56
                 :  :::  .*:  :::  :*:*::******:***::*** ::****** .* **:****

C. elegans      RGNSKIKFDEQTLELSVSSPLDRESENAIDMLIITSPPSIIHILIDVLDVNDNSPIFPID 120
C. briggsae     RGNSKIKFDEETLEFSVSSALDRESENAIDMLIVSSPPSIIHVLIDILDINDNPPKFPLE 116
                **********:***:****.*************::*******:***:**:***.* **::

C. elegans      VQRVEIPETAPIGWRVQISGATDPDEGKNGTIGKYELVDSLATVDTMSP--FGIVQSDGF 178
C. briggsae     IQNVEIPETAPIGWRVPISGATDPDQGKNGSIGKYELDEITVDGDTPTPPLFRLLQSDGF 176
                :*.************** ********:****:****** :  .  ** :*  * ::*****

C. elegans      LFLEVTGKLDRETRDLYSMRLTAIDQGVPELSSSCHLNILILDINDNPPNFGIRSLTLNW 238
C. briggsae     VYLEVVGTLDREMRDFYSMRLTASDEGVPELSASCLLNIRILDINDNPPDFGIRQIHLKW 236
                ::***.*.**** **:******* *:******:** *** *********:****.: *:*

C. elegans      NGLPNTKLFSLNATDLDSNENSLLTYRILPSGPTSEMFSISDENILVTQNNTECLQRCEF 298
C. briggsae     NGHKNAKLFLLNATDADSGDNGILKYRIQGGEEIFGILEEKDGRFLVTKNSTKCSPICEF 296
                **  *:*** ***** **.:*.:*.***  .     ::. .* .:***:*.*:*   ***

C. elegans      VVEARDSGVPPLSTTLNIVVNMEYGNEHEPNINIRFYPSDYPFIIVQPEDVNGKTLAILS 358
C. briggsae     VIEAKDSGIPPLATTLNVVVQMEYGNEHEPNINIRFYPSDFPFIVVQPEDVNGKTLAILS 356
                *:**:***:***:****:**:**********************:***:***************

C. elegans      ITDSDGPLGANSTIWIENGNEQSIFSLISRQSINILTVKHVENANQEQYILEFRANDGQS 418
C. briggsae     LTDPDGPLGPSSKIWIDSGNDQSIFSLISRQSINILTLKNVELAEKEEYTLEFAANDGQG 416
                :**.*****..*.***:.**:*****************:*:** *::*:* *** *****.

C. elegans      PADRITRKELKIFFKKYVKSTQIHVERESHVTVEKDTVPGSFVAHVETNCTDMCSFELAN 478
C. briggsae     PEEKICRRSLKIFFKKFVKSTVIQVAREIHVELERDTVAGSFVAHVETNCSEMCRFELQQ 476
                * ::* *:.*******:**** *:* ** ** :*:***.***********::** *** :

C. elegans      -SDVFKIDPFNGIIVTSSILPEGVTSYHLPIRIHLPPPSTQLVEADVFVKVIQES-VPKN 536
C. briggsae     SSDVFRIDGMNGIIVTSSELPSDVSSYHLPIRIHPPPPSTQILETDVFVKILQASSVPKN 536
                 ****:** :******** **..*:********* ******::*:*****::* * ****

C. elegans      LIRSSESPIHLKRAYTFTTWQDVSLGTVIGRLPKAQIYSTIDTVSELGVFPDGSVFVGKT 596
C. briggsae     LIRSTDPPVHLKRAYTFDTWQNVTVGTVVGKLPKAQIYSTRDLESELGVFPDGSIFVGKS 596
                ****::.*:******** ***:*::***:*:********* *  **********:****:

C. elegans      ITSDFVTLPVTLVNRNTTQTSIITLIVKPLNQHSPICQITEIHVLENAPIGTIFGRIQAR 656
C. briggsae     ISGDLVRLEVILENRNTTQTAVVTVLLKPLNRHSPVCADTKVRVLESEKPGNFIGKLHAT 656
                *:.*:* * * * *******:::*::****:***:*  *:::***.   *.::*:::*
```

Figure 4: Seqcomp and Family Relations predictions for *egl-17*, *zmp-1* and *cdh-3* upstream sequences

In these analyses the window size is 20 bp. After the Seqcomp program found a region of similarity, this region was examined by eye for other conservation near by. These regions are shown in red. In all four analyses, the translational start site is located on the far right and side of the schematics. (A) In the EGL-17 upstream comparison, we used a threshold value of 90% similarity. Elements A, B, C and D are shown on the schematic of the upstream sequence. The four smaller panels below show the nucleotide conservation of these four elements between the two species. (B) For the ZMP-1 upstream comparison, we used a 85% threshold level. (C) In the first *cdh-3* comparison, we used sequences that corresponded to sequences that resided within *C. elegans* construct mk96-134. We used a threshold of 85% identity. (D) In the second *cdh-3* comparison, we used sequences that corresponded to sequences residing within *C. elegans* construct mk96-134. We used a threshold of 85% identity.

Figure 4: Seqcomp and Family Relations predictions for *egl-17, zmp-1* and *cdh-3*

(A) *egl-17*



egl-17 element A

C. elegans TGCCTCGCCTCATCGAATTATTGGATGAGGTCAGTTATGGC
C. briggsae CCAGCAAACTCATCAAACTATTGGATGAGGTCAGTTAGTAT

egl-17 element B

C. elegans CTCCTTTTCTTTCACGTCCTTGGTAATTTTCATATGTATGTT
C. briggsae TTCTTCTTCCTTCACGTCCTTGGTAATTTTCATATGTAGAGG

egl-17 element C

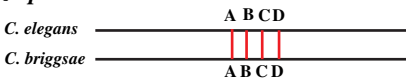
C. elegans CATCTTAATATGATGTCCAGTCAATAGTTTTCCTCAG
C. briggsae ATGAGTATAATGATGCGCAGTCAATAGTTATTCTTTC

egl-17 element D

C. elegans ACAATTGTCCGACAACTTCAAGTTGTAATTACAATGTGTTTTGAAAGAAAAAAGT
C. briggsae ACAATTGCCCGACGTCTTCAAGTTGTAATTACAATGTGTTTTGACGAAAATAAA

C. elegans AAAAAAGTGACAAAAAAGTTGATTAAATTCTTGTTTCTGATCTGATTTC
C. briggsae AAAAAAGTGATGAAAAAGTTGATTAAATTCTTGCGTCTGATTCTTTTTC

(B) *zmp-1*



zmp-1 element A

C. elegans ATGCGCCCCTCGAGAGAAGATGTATTTCGTAACCCATTTCAAAAGAAGGACGGCTCGTTGAACAG
C. briggsae TGGTGTCCCTCGAGAGAGACTCCTTCTATTTCGTAACCCATTTCGTAAGTATCGGCTCGTTGAACG

zmp-1 element B

C. elegans AAGTATTCGAGTACGTTTACACTGGTTCTG
C. briggsae CGAATCTCGAGTACGTTTACACTTTGGTTT

zmp-1 element C

C. elegans AGATGCAAACTGATTCATGTGTACGTATTGCTTGAAAAAAAGA
C. briggsae TCTCCAGAACTGATTCATGTGTGTGTGTTTGCTTATTGAAC

zmp-1 element D

C. elegans GTAGAAGGGTATTAGTCGTAGTAGTAGTATTCAGT
C. briggsae GCCAGTTTTACTACTACCAATACTAATACTACCTC

(C) *cdh-3* mk96-134 homologous region

cdh-3 element A

C. elegans GTTCAGATTCCCAAACAGAAAAAAAAACAAATAAAAAGGCAC
C. briggsae TGGCTCCATTTTCTTCTATGTTTTTTTTTCTCTTTTGTTTTC

cdh-3 element B

C. elegans TAGCCAATGTTTATGTGTCATGAATAATGAATGGTTTGGA
C. briggsae TCTAATGTTTACGTATGTTGTGAATAATGAATGGGTGGTT

cdh-3 element C

C. elegans TTTCCTGTTAATTTTATTTGACGACAACTTAATGAAAT
C. briggsae AGTTATCAAAAAGTTGTCGACTACTAAAATGTAGTTCG

cdh-3 element D

C. elegans GAAATTTGAAATGTAGCTACCAAAAAATTGCCTTGTTCTGAA
C. briggsae TACACATTTTGTTAGTGATCAATTTTTTTGTTAGCTAATTTGC

cdh-3 element E

C. elegans TCCCTATACAAAACGGACCGACCGTCCCAAAA
C. briggsae CTAACTTACAAAACGTACCGCGCGTCGTGTAT

cdh-3 element F

C. elegans CACTACCACCTGCCTTTTGTGTGTTCGTTCGCGCGTGTCCCGCCTGT
C. briggsae CCTAGCCACCTGTCTTTTTGTGTGTTCGTTCGCGCGTATCCACCATTC

(D) *cdh-3* mk66-67 homologous region

cdh-3 element H

C. elegans GATTCTCTTGTATATCCAATTTCGAGATTGTCCTTCACACCACACAGTGCCAATTGTCTTTCC
C. briggsae GTTCCTCTCATATATCCAATTTTCGAGATTGTCCTCCTCCAAAACAGTGCCAATTGTCTTTCG

cdh-3 element I

C. elegans CGAGATTGTCCTTCACACCACACAGTGCCA
C. briggsae GGCCTGTGTACTTCACACCTCACAGCCAAT

cdh-3 element J

C. elegans GGGCGGTCTCTTTCTGTTCCTCTCATAGTTCACACCTTTT
C. briggsae CCTCTTCCTCTTCCTGTTCCTCTCATATATCCAATTTTCG

cdh-3 element K

C. elegans CCAATCCAATATGTCCTTTTGATATGCTAATTTGCATTCTCTGTCCGCG
C. briggsae TTCCGCCAATATGTCCTTTTGATATGCTAATTTTCTTTTCCTCTTCTT

Table 1: Summary of construct expression patterns

This table lists the origin of the upstream region. The names of the construct, features of this construct (e.g., conserved elements (elem.) contained within the region), and the promoter from which expression is driven are listed, as well as which species was injected, and the resulting expression pattern. * This construct showed variable expression in the presumptive vulE and vulF cells, as well as variable expression in the secondary lineages, the presumptive vulA-D. Constructs mk84-148, mk50-51, mk96-134 and mk66-67 were generated in a sufficiency analysis of these three genes in *C. elegans* (Kirouac and Sternberg, in prep.).

Table 1: Summary of construct expression patterns

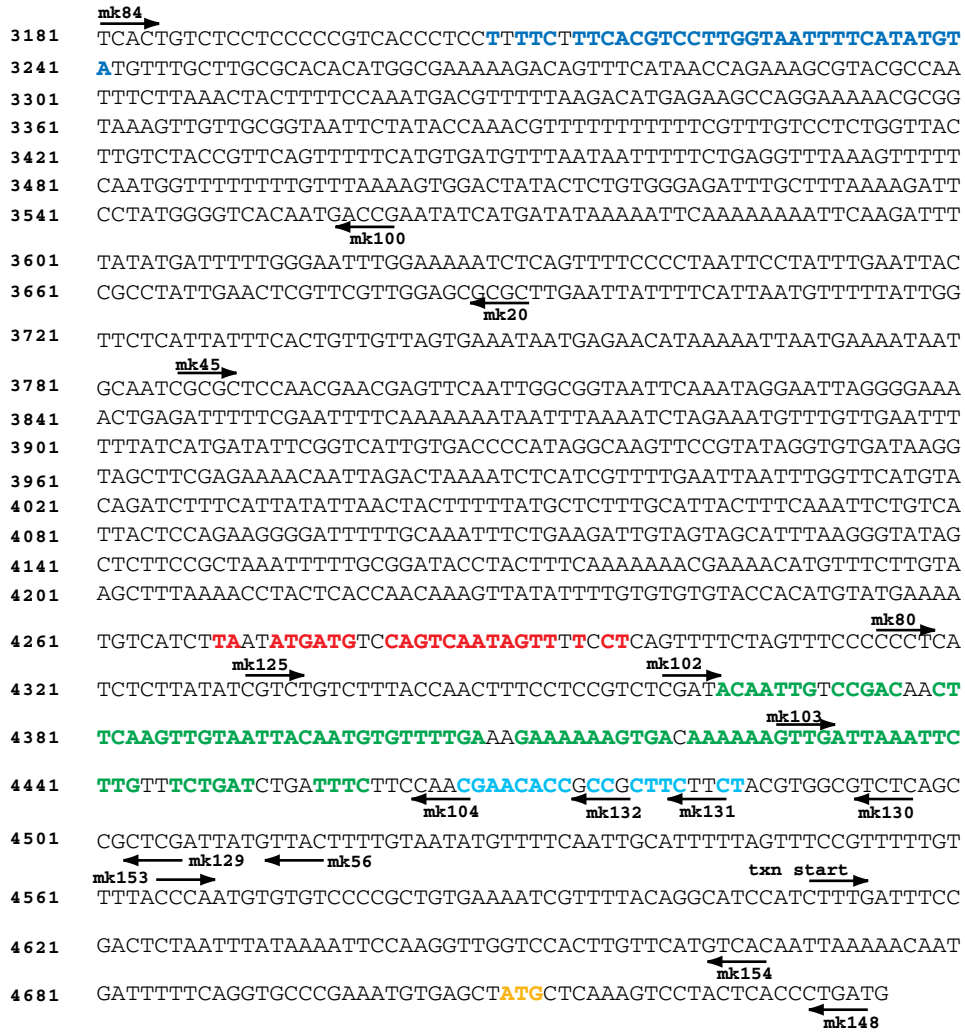
Origin	Construct Name	Features	Promoter	Species injected	Expression
Ce-egl-17	NH#293	Full length	native	*C. elegans*	Early, vulC and vulD
Ce-egl-17	NH#293	Full length	native	*C. briggsae*	Early, vulC and vulD (vulC/D slightly variable)
Cb-egl-17	mk160-161	Elem. B-E	*pes-10*	*C. elegans*	Variable early*, vulC and vulD
Cb-egl-17	mk160-161	Elem. B-E	*pes-10*	*C. briggsae*	No early, variable vulC and vulD
Ce-egl-17	mk84-148	Elem. B-E	*pes-10*	*C. elegans*	Early, vulC and vulD
Ce-zmp-1	pJB100	Full length	native	*C. elegans*	vulE, vulA and anchor cell
Ce-zmp-1	pJB100	Full length	native	*C. briggsae*	vulE, vulA and anchor cell
Cb-zmp-1	mk172-173	Elem. A-D	*pes-10*	*C. elegans*	vulE, vulA and anchor cell
Ce-zmp-1	mk50-51	Elem. A-D	*pes-10*	*C. elegans*	vulE, vulA and anchor cell
Ce-cdh-3	jp#38	Full length	native	*C. elegans*	vulE, F, C and D and anchor cell
Ce-cdh-3	jp#38	Full length	native	*C. briggsae*	anchor cell, rare vulval cell expresses
Cb-cdh-3	mk162-163	Elem. A, B, and D-F	*pes-10*	*C. elegans*	vulE, F, C, and D
Ce-cdh-3	mk96-134	Elem. A-F	*pes-10*	*C. elegans*	vulE, F, C and D and anchor cell
Cb-cdh-3	mk164-165	Elem. H-K	*pes-10*	*C. elegans*	vulE, F, C (variable) not vulD
Ce-cdh-3	mk66-67	Elem. H-K	*pes-10*	*C. elegans*	vulE, F, C and D

Figure 5: *egl-17* nucleotide sequences of important regions

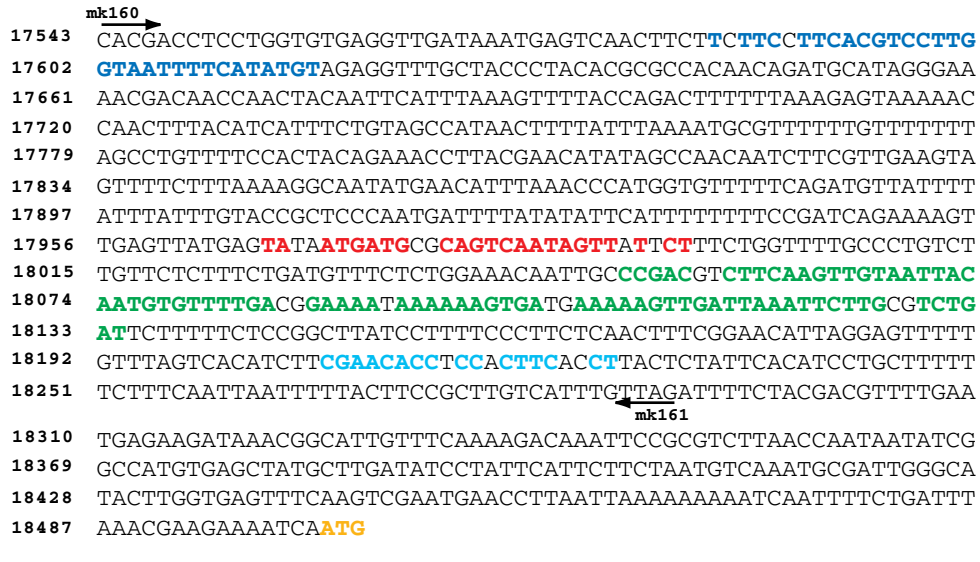
(A) The nucleotide sequence of *C. elegans egl-17* mk84-148 is shown. The *egl-17* genomic region of NH#293 contains 3819 bp of upstream sequence. The first exon of the transcript starts at nucleotide 4610, and translation starts at nucleotide 4708. Nucleotide 790 of the *egl-17* upstream region corresponds with nucleotide 17648 in Genbank cosmid F38G1 (Accession # AC006635). (B) The nucleotide sequence of *C. briggsae egl-17* upstream region mk160-translational start site is shown. The *C. briggsae egl-17* upstream region lies on contig c000300114 (nucleotides 17543-18504). Arrows show the end points and direction of primers in the region. The conserved elements found by the Seqcomp and Family relations programs are depicted in different colors. Note that neither of these sequences shows conserved element A.

Figure 5: *egl-17* nucleotide sequences of important regions

(A) *C. elegans egl-17* mk84-148



(B) *C. briggsae egl-17* mk160-translational start site



element B
element c
element D
element E
translational start site of EGL-17
indicate boundaries of primers

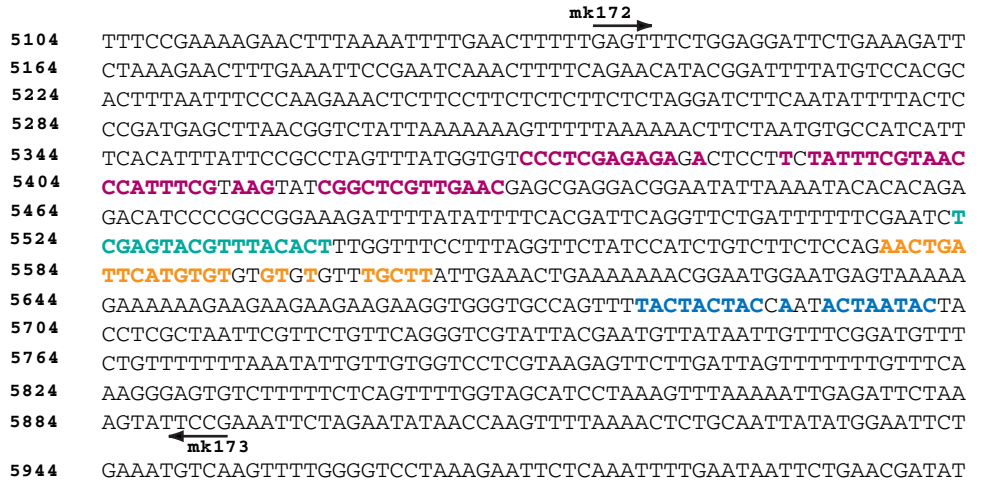
Figure 6: *zmp-1* nucleotide sequences of important regions

(A) The nucleotide sequence of *C. elegans zmp-1* mk50-51 is shown. The *zmp-1* genomic region in pJB100 contains 3472 bp of upstream sequence. The translational start site of ZMP-1 is at nucleotide 3473. Nucleotide 1 of this *zmp-1* upstream region corresponds with nucleotide 7630 in Genbank cosmid EGAP1 (Accession # U41266). In this panel, nucleotides 992-1438 are shown. (B) The *C. briggsae zmp-1* upstream region mk172-173 that contains the conserved elements predicted by Seqcomp program lies on contig c010400937. Arrows show the end points and direction of primers in the region. The conserved elements found by the Seqcomp and Family relations programs are depicted in different colors.

Figure 6: *zmp-1* nucleotide sequences

(A) *C. elegans zmp-1* mk50-51

(B) *C. briggsae zmp-1* mk172-173



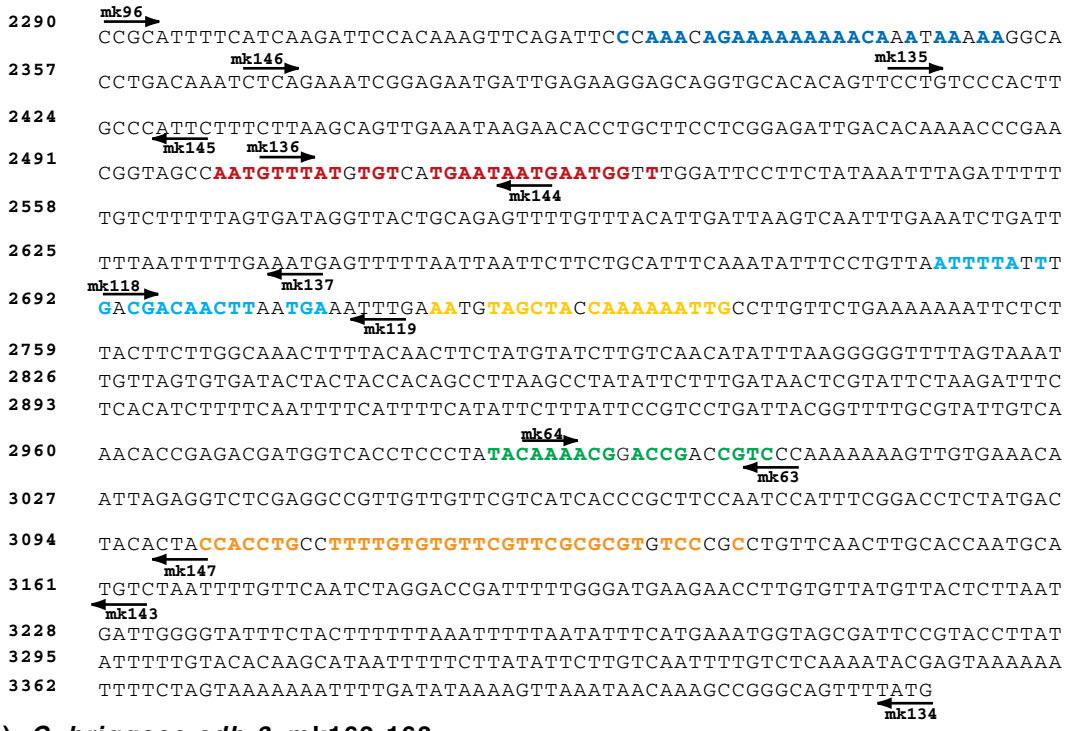
element A
element B
element C
element D
indicate primers location

Figure 7: *cdh-3* nucleotide sequences of mk96-134 and mk162-163

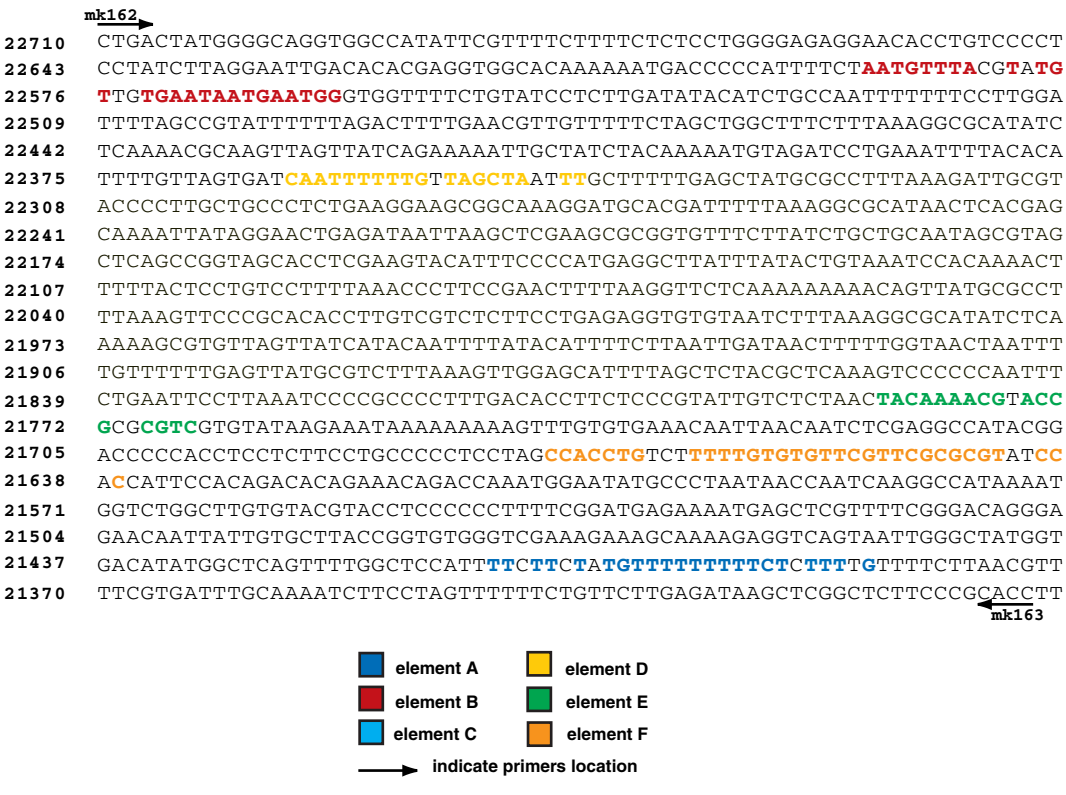
(A) The nucleotide sequence of *C. elegans cdh-3* mk96-134 is shown. The jp#38 genomic region of *cdh-3* contains 5928 bp of upstream sequence, whose start codon occurs at nucleotide 6041. Nucleotide 113 of the *cdh-3* upstream region corresponds with nucleotide 37343 in Genbank cosmid ZK112 (Accession # L14324). In this panel, nucleotides 2290-3419 are shown. (B) The *C. briggsae cdh-3* upstream region mk162-163 that contains conserved elements predicted by Seqcomp program lies on contig c014100642 (20582-22703). Arrows show the end points and direction of primers in the region. The conserved elements found by the Seqcomp and Family relations programs are depicted in different colors. Note that elements C and E that are found in *C. elegans* mk96-134 are not in mk162-163.

Figure 7: *cdh-3* nucleotide sequences of mk96-134 and mk162-163

(A) *C. elegans cdh-3* mk96-134



(B) *C. briggsae cdh-3* mk162-163



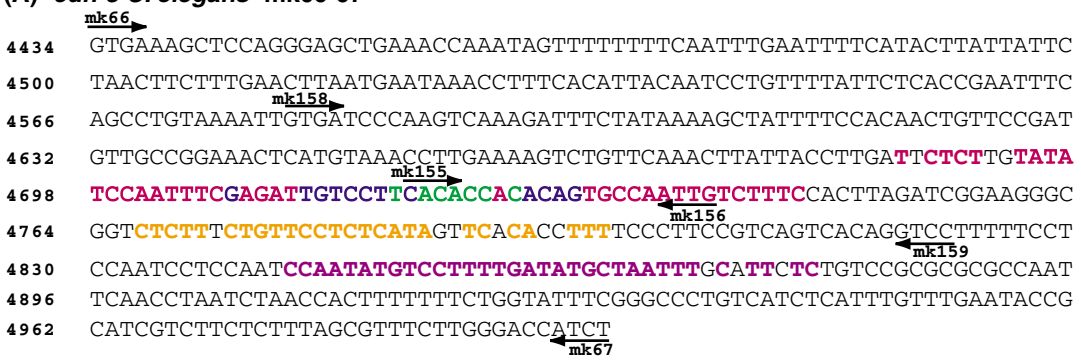
element A element D
element B element E
element C element F
indicate primers location

Figure 8: *cdh-3* nucleotide sequences of mk66-67 and mk164-165

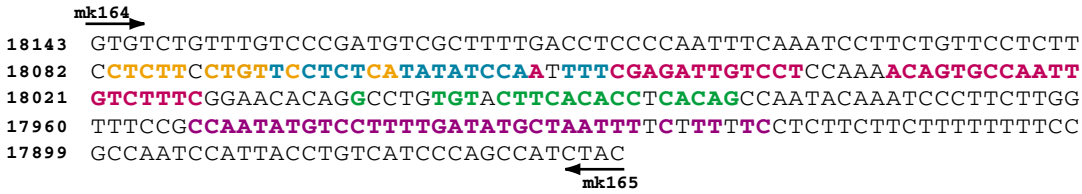
(A) The nucleotide sequence of *C. elegans cdh-3* mk66-67 is shown. The jp#38 genomic region of *cdh-3* contains 5928 bp of upstream sequence, whose start codon occurs at nucleotide 6041. Nucleotide 113 of the *cdh-3* upstream region corresponds with nucleotide 37343 in Genbank cosmid ZK112 (Accession # L14324). In this panel, nucleotides 4434-4997 are shown. (B) The *C. briggsae cdh-3* upstream region, mk164-165, which contains conserved elements predicted by the Seqcomp program lies on contig c014100642 (nucleotides 17869-18145). Arrows show the end points and direction of primers in the region. The conserved elements found by the Seqcomp and Family relations programs are depicted in different colors. Note that elements H and J overlap in mk164-165, and elements H and I overlap in mk66-67.

Figure 8: *cdh-3* nucleotide sequence mk66-67 and mk164-165

(A) *cdh-3 C. elegans* mk66-67



(B) *cdh-3 C. briggsae* mk164-165



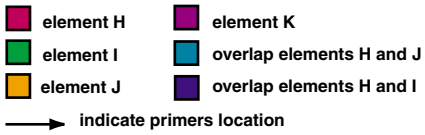
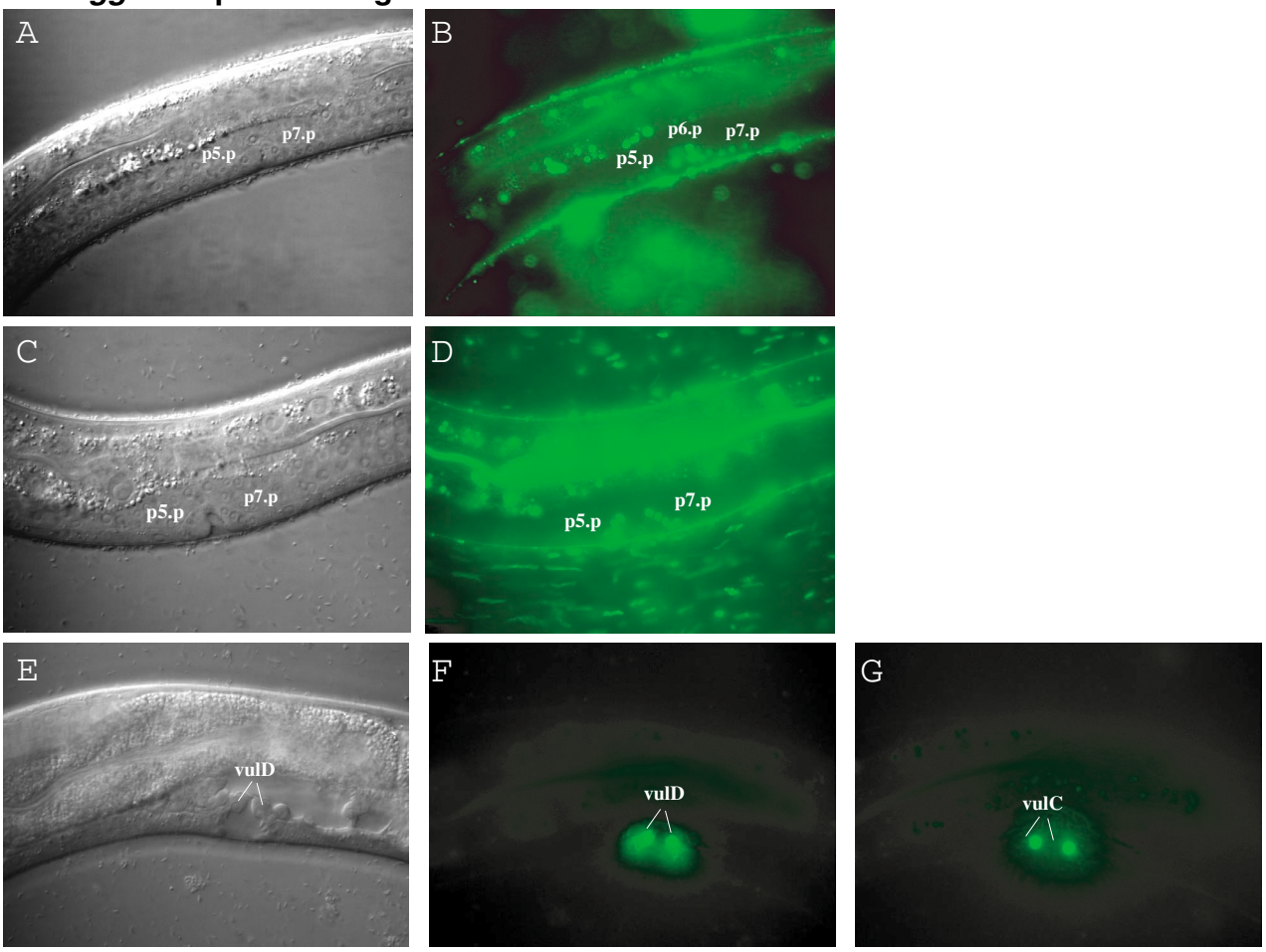


Figure 9: *C. briggsae* upstream regions injected in *C. elegans*

Panel A shows the expression pattern of *C. briggsae* mk160-161 when it is injected into *C. elegans*. mk160-161 (A) Nomarski DIC photomicrograph of an animal as the vulva has started to invaginate is shown. mk160-161 (B) All of P5.p, P6.p, and P7.p are GFP positive. Another example of this variable expression pattern is seen images C and D. mk160-161 (C) Nomarski DIC photomicrograph of a slightly older animal. mk160-161 (D) The fluorescent image of this same animal is seen; clear expression is seen in the descendants of P5.p and P7.p, but not in P6.p (not in this focal plane and not expressing). mk160-161 (E) Nomarski DIC photomicrograph of an L4 animal with vulD cells labeled. The vulC cells are not in this plane of focus. mk160-161 (F) This is the same animal and the fluorescence is clearly visible in vulD cells. mk160-161 (G) The same animal is shown again in a slightly different focal plane to see the GFP expression in the vulC cells. In panel (B), are shown some representative pictures from *C. elegans* animals that were injected with *C. briggsae* mk162-163. mk162-163 (A) Nomarski DIC photomicrograph of an animal that has just start to invaginate. The P6.p, the presumptive vulE and vulF, cells are labeled. mk162-163 (B) Shows the fluorescence image of the same animal and GFP is clearly seen in both vulE and vulF cells. mk162-163 (C) Nomarski DIC photomicrograph of an L4 animal, with vulD cells labeled. The vulC cells are not in this plane of focus. mk162-163 (D) Same animal; fluorescence is clearly visible in vulD cells. mk162-163 (E) Same animal again in a slightly different focal plane. The GFP in vulC cells is evident. All photomicrographs are lateral views of the animals.

Figure 9: *C. briggsae* upstream regions

(A)



(B)

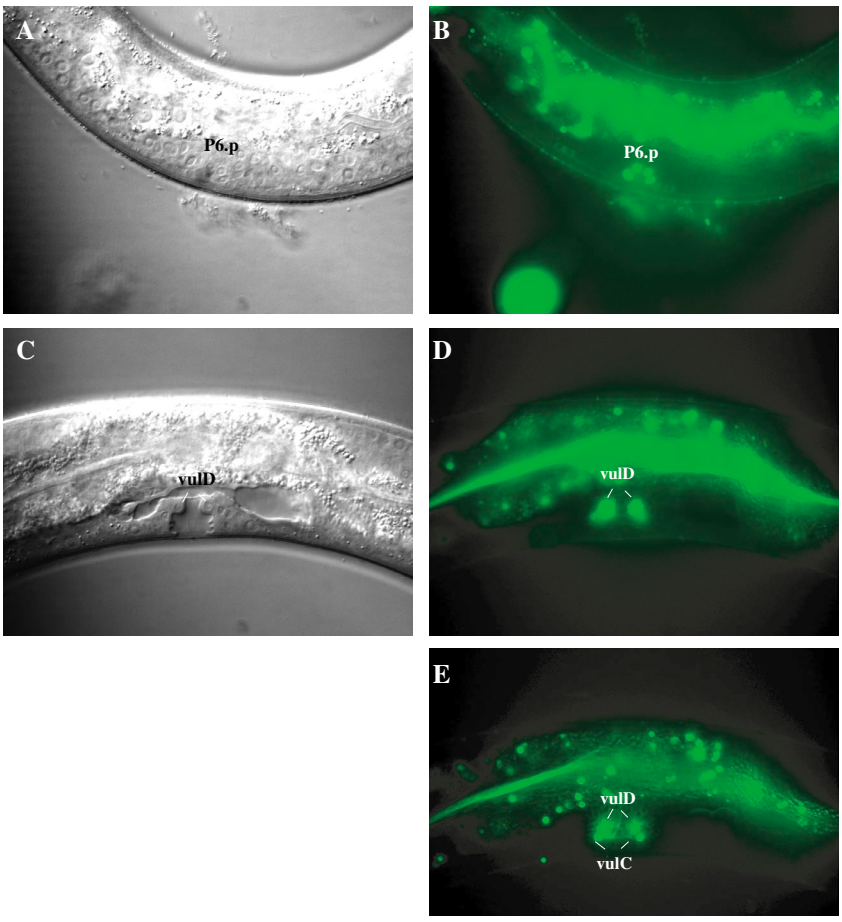


Table 2: Transfac binding site predictions in regions of similarity between *C. elegans* and *C. briggsae*

Transfac prediction binding sites were listed that meet the following criteria: (1) the minimum core binding specificity had to be at least 0.90, (2) the maximum Random Expectation Value, "re", which is the number of times this site would appear in a random 1000 bp, was not exceed 0.51, and (3) the sites had to appear in both the *C. elegans* region and the homologous *C. briggsae* region. The number of sites in the *C. elegans* region is followed by a slash, and then the number of sites in the *C. briggsae* region is listed. In addition, if the site was in a conserved region, inside the parentheses is denoted how many sites are conserved and in what element. There are several factors marked by *: these factors where not necessarily found in both *C. elegans* and *C. briggsae*, but were included because they are part of some potentially interesting transcription families. The letters B, C, D, F and K refer to the conserved elements in these regions.

Table 2: Transfac database prediction in conserved regions

		egl-17	*zmp-1*	*cdh-3*	*cdh-3*
FAMILY OF FACTORS	FACTOR	mk84-148/ mk160-161	mk50-51/ mk172-173	mk96-134/ mk162-63	mk66-67/ mk164-165
AP1 and related factors	NFE2.01		1/1 (1, B)		
Arabidopsis HomeoBox Protein	ATHB1.01	6/1			
ARS binding factor	ABF1.01		1/1	1/2	
ARS binding factor	ABF1.01		1/		
ARS binding factor	ABF1.02		2/2		
Aspergillus Spore/Developmental regulator	ABAA.01		1/2		
Brn POU domain factors	BRN3.01	5/1			
C. elegans maternal gene product SKN-1	SKN1.01	2/1			1/1
cAMP-Responsive Element Binding proteins	E4BP4.01		2/2 (1, D)		
Ccaat/Enhancer Binding Protein	CEBP.02		1/1 (1, D)		
Cell-death specification 2	CES2.01		2/2 (1, D)		
CLOX FAMILY	CDP.01	1/1	1/1	1/3 (1, B)	
CLOX FAMILY	CDPCR3.01	3/2			1/2 (1, K)
CRP binding Site	CRP.01			1/2	
BRoad-Complex ecdysone steroid response	BRCZ4.01			1/1	
Drosophila gap gene hunchback	HB.02	4/2 (1, D)		1/3	
E2F-myc activator/cell cycle regulator	E2F.01			2/3	
E2F-myc activator/cell cycle regulator	E2F.03			2/2 (1, F)	
ETS	c-ETS-1 (p54) *	0/1			0/1
ETS	ETS1.01			2/1	
ETS	PU.1ETS *	2/1		0/1	1/1
EVI myleoid transforming protein	EVI1.01			1/2	
EVI myleoid transforming protein	EVI1.02		1/2	2/2	
Floral determination	MADSA.01		1/1	4/7	
Fork Head and Related	FREAC2.01	1/3		2/2	
Fork Head and Related	FREAC4.01		1/1		
Fork Head and Related	XFD2.02			1/1	
GATA FAMILY	GATA1.04			1/1	
Glucocorticoid Responsive	ARE.01			2/1	
Glucocorticoid Responsive	GRE.01				1/1 (1, K)
Glucocorticoid Responsive	PRE.01		1/1	1/2	
Homeodomain Factor	FTZ.01	4/1		3/6	
Homeodomain Factor	NKX25.02		1/2		
Homeodomain Factor	NKX31.01		1/1		

Homeodomain Factor	PBX1.01			3/2	
Homeodomain Factor myeloid lukemia	MEIS1.01	3/2	1/2 (1, B)		
Homeodomain Pancreatic /Intestinal LIM domain	ISLI1.01	5/2 (1, D)	1/1	2/3	2/1 (1, K)
Homeodomain Pancreatic /Intestinal	PDX.01	1/1 (1, D)			
Homeoprotein Caudel	CDX2.01	5/2	1/4	4/3	
HOXF	HOX1-3.01	5/2		1/4 (1, B)	
HOXF	HOXA9.01		1/1 (1, B)		
HSF family	FHSF.01		1/3		2/1
HSF family	FHSF.02		1/1		
HSF family	FHSF.03	2/2	1/5	2/1	
HSF family	FHSF.04		1/2		
HSF family	IHSF.01		1/2		
HSF family	IHSF.03	2/1			
HSF family	IHSF.04		1/3		
Interferon Regulated Factor	IRF1.01		1/2	2/2	1/1
Interferon Regulated Factor	IRF2.01		1/2		
Interferon Regulated Factor	ISRE.01		1/1		
MEF2-myocyte-specific enhancer-binding	AMEF2.01	1/1		2/6	
MEF2-myocyte-specific enhancer-binding	HMEF2.01			1/2	
MYB-Like protein (Petunia)	MYBPH3.01		1/2 (1, D)	1/7	
Octamer Family	OCT1.01	1/1	1/1		
Octamer Family	OCT1.06	6/2		1/2	
Octamer Family	OCT1.06	4/2 (1, D)			
papillioma virus E2 Txn activator	E2.02			1/1	
PAX3 FAMILY	PAX3.01		1/1 (1, D)		
Phaseolus vulg. SILencer reg. of chalcone	SBF1.01	3/3	1/3	3/6	
Plant I-Box sites	IBOX.01	1/1			
Plant P-Box binding sites	PBOX.01			1/2	
Poly A	APOLYA.01	3/3			
Poly A	POLYA.01	1/1		1/1	
Promoter-CcAaT binding	ACAAT.01		1/1	1/1	2/2
Repr. of RXR-mediated activ. & retinoic	COUP.01			1/1	
signal transducers and activators of txn	ISTAT.01			1/1	2/2
signal transducers and activators of txn	STAT6.01	1/1			
SMAD Family TGF-B	FAST1.01		2/2	1/2	
Special AT rich binding Sequence	SATB1.01		1/1		
TATA FAMILY	TATA.02	6/2			
Tata-Binding Protein Factor	ATATA.01	2/1		2/5	
TCF/LEF	LEF1.01 *	1/1	1/3	2/2	
TCF/LEF T-cell Homolog	TCF/LEF *			0/1	

TCF/LEF	TCF/LEF *	2/1		0/1	1/1
TCF/LEF	TCF11/KCR-F1/NRF1 *	2/1		1/1	
Vertebrate steroidogenic	SF1.01			1/1	
XhoI site-binding protein I	XBP1.01	1/1			
Yeast CCAAT binding	HAP234.01	4/3		1/2	
Yeast GC-Box Proteins	MIG1.01			1/1	
Yeast MADS-Box factors	RLM1.01			1/1	
zinc finger W Box family	WRKY.01	1/3 (1, C)		2/1	1/1
zinc finger *Xenopus* MYT1 C2HC	MYT1.01	1/1			
zinc finger *Xenopus* MYT1 C2HC	MYT1.02	5/6 (1, D)	1/9	4/8	

Table 3: AlignACE predictions of overrepresented sequences

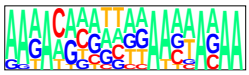
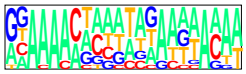
(A) A summary of the number of motifs found in each of the listed regions. The total number of motifs identified by AlignACE is shown in parentheses, while the number of motifs that scored above the MAP score threshold of ten is shown outside the parentheses for both the eight- and 10-bp motifs. The last entry on this table is a comparison of *C. elegans cdh-3* to *C. briggsae zmp-1*, each of which drives expression in the anchor cell. As indicated in the left-hand column, this comparison was performed to isolate motifs that might be important in conferring anchor cell expression on a naïve promoter. (B) This table summarizes the data for each of the motifs listed in Table 3A that had a MAP score over 10. The region is listed in the left-hand column. The motif numbers are consecutive and are followed by the size of the motif. The MAP score for each motif is shown under the column head MAP. The sites for each motif are listed. If more then one region was being compared, the sites for the first as indicated by the left-hand column are in parentheses, followed by the second set of parentheses, and so on. Abbreviations are as follows: expr. stands for expression; imp. stands for importance and elem. stands for element. The pictograms were generated using the Pictogram program (http://genes.mit.edu/pictogram.html).

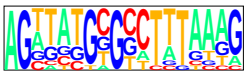
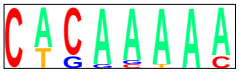
Table 3: AlignACE predictions of overrepresented sequences

A.

Expression	Regions examined	Gene	8 bp motif	10 bp motif
	mk160-161	*Cb-egl-17*	3 (7)	2 (7)
	mk172-173	*Cb-zmp-1*	3 (6)	2 (6)
	mk162-163	*Cb-cdh-3*	9 (12)	5 (7)
	mk164-165	*Cb-cdh-3*	1 (4)	2 (3)
Anchor cell	mk96-134/mk172-173	*Ce-cdh-3/Cb-zmp-1*	1 (12)	2 (13)

B.

Region	Motif	MAP	Consensus	Sites	Comments
mk160-161 *Cb-egl-17*	1.8	13.77		37, 206, 224, 296, 350, 378, 391, 447, 474, 546, 559, 591, 645, 705, 721	Site 37 is located in conserved element B imp. for early expr. Site 559 is located in conserved element D imp. for vulC/D expr.
	2.8	13.33		37,116,206, 226, 378, 453, 474, 521, 548, 591, 644, 704	Site 37 is located in conserved element B imp. for early expr. Sites 521 and 548 located in conserved element D imp. for vulC/D expr.
	3.8	10.84		37, 112, 168, 224, 293, 388, 446, 469, 542, 588, 609, 645, 700, 717	Site 37 located in conserved elem. B imp. for early expr. Sites 542 and part of 588 located in conserved element D imp. for vulC/D expr.
	4.10	11.96		62, 114, 206, 226, 280, 320, 344, 378, 443, 471, 521, 549, 579, 705	Sites 549 and 579 are located in conserved element D imp. for vulC/D expr. Multiple sites overlap motif 2.8 sites.
	5.10	10.19		38, 112, 168, 293, 307, 331, 388, 456, 469, 483, 592, 609, 700, 717	Site 38 is located in conserved element B imp. for early expr. Multiple sites overlap motif 3.8.
mk172-173 *Cb-zmp-1*	1.8	13.64		12, 22, 42, 122, 278, 361, 376, 411, 472, 584, 614, 716, 738	Site 278 is located in conserved element D. This is the only one of motifs in this element that is present in mk50-51
	2.8	13.46		30, 48, 62, 112, 131, 183, 367, 425, 480, 505, 575, 605, 624, 669, 700, 730	
	3.8	10.69		5, 22, 112, 249, 270, 352, 373, 408, 472, 497, 513, 575, 615, 673, 708, 738	Sites 249 and 270 are located in conserved element D.
	4.10	16.17		5, 22, 119, 270, 352, 373, 408, 472, 497, 513, 575, 615, 673, 708, 738	Site 270 is located in conserved element D. Multiple sites overlap 3.8 motif sites.
	5.10	13.21		1, 21, 94, 119, 174, 373, 408, 546, 705, 737	Site 546 is located in conserved element A. Multiple Sites overlap motif 1.8 sites.

Region	Motif	MAP	Consensus	Sites	Comments
mk162-163 *Cb-cdh-3*	1.8	23.92		232, 249, 265, 368, 384, 443, 495, 520, 549, 665, 715, 731, 804, 820	
	2.8	23.66		8, 45, 57, 610, 678, 887, 903, 1006, 1021, 1158, 1196, 1226, 1397	
	3.8	22.62		6, 56, 256, 381, 410, 450, 611, 662, 722, 888, 1022, 1108, 1159, 1194, 1249, 1278	
	4.8	20.83		259, 376, 396, 453, 489, 528, 657, 725, 812, 1062, 1103, 1151, 1217, 1379	Site 1062 is located in conserved element F imp. for vulval expression and in anchor cell gamma region.
	5.8	19.60		251, 377, 445, 502, 585, 658, 717, 813, 878, 997, 1026, 1256, 1378	Site 1028 is located in conserved element F imp. for vulval expression and in anchor cell gamma region.
	6.8	15.75		7, 44, 56, 103, 559, 673, 855, 882, 901, 1001, 1020, 1157	
	7.8	13.39		96, 154, 267, 288, 371, 594, 645, 733, 807, 968, 1045, 1085, 1212, 1338, 1366	Site 1045 is located in conserved element F imp. for vulval expression and in anchor cell gamma region.
	8.8	12.84		20, 180, 239, 270, 362 421, 567, 593, 651, 736, 962 1010, 1036, 1177, 1239, 1302, 1319, 1357 Site 362 in conserved elem. D. Site 1036 in conserved elem. F and Site 1302 and 1319 in conserved elem. A imp region for vulA cell expr.	
	9.8	10.63		98, 150, 251, 383, 438, 631, 664, 717, 878, 1087, 1131, 1162, 1186, 1255, 1281	
	10.10	25.99		258, 376, 452, 657, 724, 812, 1061, 1379	Site 1061 is located in conserved element F imp. for vulval expression and in anchor cell gamma region.
	11.10	21.47		6, 43, 55, 397, 407, 608, 885, 1006, 1019, 1156, 1194	

Region	Motif	MAP	Consensus	Sites	Comments
	12.10	19.16		20, 180, 239, 270, 362, 736, 798, 958, 1036, 1177, 1239, 1302, 1319, 1357	All sites, except 798 and 958 are the same as motif 8.8. See 8.8 comments.
mk162-163 *Cb-cdh-3*	13.10	16.61		26, 94, 232, 263, 301, 370, 641, 686, 729, 768, 786, 806, 952, 1044, 1189, 1239, 1296, 1320, 1336, 1365 Site 351 in conserved elem. D, Site 1044 in conserved elem. F, and Sites 1296 and 1239 in conserved elem. A	
	14.10	10.43		79, 251, 382, 445, 663, 717, 818, 1228	Overlaps multiple sites with motif 9.8.
	1.8	15.29		1, 26, 52, 70, 99, 121, 143, 183, 195, 221, 239, 258	Sites 99 and 121 are in conserved element H. Site 143 is located in conserved element I and Site 194 is located in conserved element K.
mk164-165 *Cb-cdh-3*	1.10	17.49		3, 19. 45, 63, 92, 176, 214, 232	Site 45 is located in the overlap elements J and H. Site 63 is located in element. H.
	2.10	11.02		19, 45, 63, 92, 141, 176, 214, 232	Site 63 is located overlap elements J and H. Site 92 located in element H. Site 141 located in element I. Site 214 located in element K.
	1.8	16.88		[38, 50, 81, 248, 263, 333, 374, 463, 484, 591, 610, 723, 820, 948, 964, 1027] [345, 367, 473, 501, 513, 623, 698, 722]	Sites 38 and 50 in mk96-134 are located in conserved element A of *cdh-3 ,* which is imp. for anchor cell (alpha region). No sites in conserved elements of *zmp-1*.
mk96-134 *cdh-3* mk172173 *Cb-zmp-1*	2.10	16.32		[25, 45, 143, 264, 353, 384, 425, 472, 574, 719, 956, 1036, 1095] [86, 168, 373, 507, 612, 630, 662]	Site 25 in mk96-134 is partially in conserved element A of *cdh-3,* which is imp. for anchor cell (alpha region). No sites in conserved elements of *zmp-1*.
	3.10	16.01		[38,74 256, 333, 374, 416, 590, 616, 871, 948, 964, 1027] [31, 306, 345, 366, 473, 501, 623, 673, 699]	Site 38 in mk96-134 is located in conserved element A of *cdh-3,* which is imp. for anchor cell (alpha region). No sites in conserved elements of *zmp-1*. Multiple sites overlap motif 1.8 sites.