SUPPORTING INFORMATION:

Results:

Background GFP expression

Even when no element was inserted, some background expression from the pPD107.94 expression vector was observed in the posterior and anterior-most intestine, enteric muscle, anal-depressor cell, anterior-most bodywall muscle, and the anterior excretory cell (Figure S4B). Background expression varied, both in level of expression and in which cells were most strongly expressing the reporter, between different independent lines. No expression recorded in these cells expressing background was regarded as a positive hit. A second, independent reporter with a different basal promoter was also injected, pPD95.75. Its background expression patterns were the same as those observed for pPD107.94, suggesting that the Δ*pes-10* basal promoter is not affecting expression patterns. Both reporters share the same *unc-54* 3'UTR, and it may be responsible for the observed background expression.

Sequence analyses

To identify regulatory elements shared by different Hox sub-clusters, the *C. elegans*, *C. briggsae*, and *C. remanei ceh-13/lin-39* sequences were compared with their corresponding *egl-5/mab-5* sequences. We found only one similarity between all of them, corresponding to the N9 MUSSA match. While region N9 was previously known in *ceh-13/lin-39*, its presence in another sub-cluster had not been reported (see Discussion). The remaining *ceh-13/lin-39* regions should therefore be specific to that subcluster alone (Figure S9B-D).

To define genome-wide occurrences of the MUSSA-derived conserved sequences, Cistematic (Mortazavi et al. 2006) was used to scan the *C. elegans* genome for sequences that held 80% or greater similarity to the position frequency matrix (PFM; Wasserman and Sandelin 2004) generated from *C. elegans*, *C. briggsae*, *C. remanei*, and *C. brenneri* conserved sequences. The resulting hits, generally ~30-200, from the genome were then used to generate a new, refined PFM. A second round of scanning the genome using this refined PFM was used to generate a further refined PFM. Due to the AT-richness of the *C. elegans* genome using a neutral background, only CG-rich motifs survived refinement. A coherent motif identified for the N2-1 MUSSA-derived sequence was very similar when generated with searches in the *C. elegans*, *C. briggsae*, or *C. remanei* genomes (Figure S10; Mortazavi et al. 2006). Further rounds

of scanning and refinement did not change this N2-1 PFM noticeably. Such consistency through refinements and across several genomes suggests that a valid genome-wide motif may have been identified.

In the *C. elegans* genome, the refined N2-1 motif identifies 625 protein-coding genes in the WS190 release of WormBase, of which 407 had been annotated with one or more Gene Ontology (GO) terms by August 2008. These include three Hox genes: *ceh-6*, *egl-5*, and *lin-39* itself. Using GOstat (Beissbarth and Speed 2004) to determine statistically overrepresented GO terms in this N2-1 gene set, we found the three most significant terms were "small GTPase mediated signal transduction" (GO:0007264; 16 genes; *p*-value = 0.00971), "vulval development" (GO:0040025; 15 genes; *p*-value = 0.0164), and "reproductive behavior" (GO:0019098; 22 genes; *p*-value 0.0309). These are consistent with N2's expression pattern (Table 1), which includes P cells ancestral to vulval precursor cells and ventral cord motorneurons.

Since expression directed by the N3 region does not require the core N3 MUSSA match (see above), other regulatory motifs outside the core sequence must drive expression in the mutation assays and the trans-phylum assays. In addition to the N3 MUSSA match itself, MEME identified two motifs shared by the N3 regions in nematodes and vertebrates (Figure S3C). Although they have not been functionally tested, they resemble Pax4 binding sites as defined in the JASPAR database (Bailey and Elkan 1994; Sandelin et al. 2004). Moreover, the core N3 MUSSA match and an extension of it by MEME resemble LM115 and LM171 from the JASPAR CNE database of 12-22 nt motifs overrepresented in conserved, non-coding mammalian DNA (Bryne et al. 2007, Xie et al. 2007). In contrast, MEME scans of the N7 regions in nematodes and vertebrates revealed only one motif shared by these two clades, the core N7 MUSSA match (Figure S3D). Both N3 and N7 resemble the 14-nt consensus of motif LM115, with 1- or 2-nt mismatches (N7 and N3, respectively). Moreover, the subtly conserved 5'-flank of N3 has a 2-nt mismatch to motif LM171. These correlations with independently generated mammalian motifs suggest that N3 and N7 define sequences relevant to both nematode and mammalian biology. As a negative control, we used MEME to compare nematode N3 sequences to *Drosophila* Hox cluster sequences that are well-conserved in flies but not similar to worm N3; in this case, MEME only produced motifs separated strictly between these two clades (Figure S3E), suggesting that those motifs found by MEME to be shared by nematode and vertebrate N3 sequences are significant

Threshold revision

To refine our parameters, we varied the window size from 15 to 30 bp in two-, three-, four-, and five-way analyses with different combinations of *Caenorhabditis* species (Figures S2B, E-L). We recorded the maximum threshold at which MUSSA matches were observed within each of our previously defined regions (Figure S5). Averaging the maximum thresholds for two window sizes, 15 bp and 20 bp, and using a threshold of 92% had an identical yield to the 15-bp window results alone. Although these two approaches yielded the same results, the greater dynamic range observed from averaging the results may be useful when applied to other genes.

Among the novel assembled sequences of *C. brenneri* and *C.* sp 3 PS1010 were those of *lin-3*, an EGF family growth factor, and *lin-11*, a LIM homeodomain transcription factor, which both have regulatory elements known to be necessary for vulval development (Gupta and Sternberg 2002; Hwang and Sternberg 2004). We found that MUSSA matches corresponded with some, but not all, experimentally validated regulatory sites (Figure S8A, B). However, we could detect the missed sites by scanning exhaustively in the vicinities of the MUSSA matches for short overrepresented motifs with the YMF/Explanators program (Blanchette and Sinha 2001; Sinha and Tompa 2002). *C. elegans* motifs were easily found by YMF/Explanators in *C. brenneri*, but were completely missing from *C.* sp. 3 PS1010. For a 60-nt *lin-3* element active in anchor cells (Hwang and Sternberg 2004), E-box and Ftz-F1 motifs were easy to find, but their statistical significance (Z-scores) improved steadily as species number increased from two to four (Figure S8C; see Table S6). In a 460-nt element of *lin-11* driving uterine expression (Gupta and Sternberg 2002), which was larger and thus more challenging to scan for motifs, at least three genomic sequences (from *C. elegans*, *C. briggsae*, and *C. remanei*) were required to detect the crucial LAG-1 binding motifs (Figure S8D). None of the ACEL or LAG-2 motifs were found in *C.* sp. 3 PS1010's *lin-3* or *lin-11* genes. If the 5' region of *C.* sp. 3 PS1010's *lin-3* was included in a motif scan, Z-scores fell by two-thirds; including the *lin-11* 5' region had less dramatic but still visible detrimental effects (Table S6). Moreover, while the regulatory elements in the Elegans group species were associated with several motifs, *C.* sp. 3 PS1010's genes lacked such groups of motifs (Figure S8). We scanned contig sequences surrounding *C.* sp. 3 PS1010 *lin-3* and *lin-11* (~30 kb in each direction) in case these elements might exist at a greater distance from their genes, but this yielded no MUSSA matches or motif clusters. These examples also show that inclusion of sequences from a divergent worm genome (*C.* sp. 3 PS1010)

can lower the success rate for finding validated elements, as in *ceh-13/lin-39*. *lin-3* and *lin-11* also illustrate complementary computational approaches: MUSSA can collect regions in additional genomes for refined input to motif search algorithms, which in turn are more successful than they would have been with unrefined inputs.

Author contributions

SGK, EMS, BJW, and PWS conceived and designed the experiments. TDB and DT designed and wrote the MUSSA software. JAD and HS prepared and sequenced the *C. brenneri* and PS1010 clones. EMS merged raw sequence assemblies, annotated them, ran the comparative analysis for the *lin-3* and *lin-11* genes, and identified exotic Hox clusters and JASPAR CNE motifs. SGK ran comparative analyses, performed the *in vivo* experiments, and analyzed the resulting data for the *ceh-13/lin-39* Hox cluster and non-nematode Hox clusters. SGK, EMS, BJW, and PWS wrote the paper.

Methods

General methods and strains. Genomic DNA used as carrier in microinjections was digested 5-fold with XbaI, HinDIII, NcoI, XhoI, EcoRI, and BamHI (New England Biolabs) and phenol-chloroform purified. At least three independent and stable transgenic lines were generated for each construct. Negative controls, including the digested genomic DNA, gave no GFP expression except for the expected background from controls with pBluescript. Mosaic animals were utilized for expression studies.

Strain and culture conditions. *Caenorhabditis brenneri* was first isolated as a single strain (CB5161) from sugar cane in Trinidad by D.J. Hunt (Sudhaus and Kiontke 1996). Unlike *C. elegans* and *C. briggsae*, but like most other nematode species, *C. brenneri* is gonochoristic, with male and female sexes rather than males and hermaphrodites (Kiontke et al. 2004). *Caenorhabditis* sp. 3 PS1010 was first isolated as a single strain, PS1010 (Baldwin et al. 1997), and like *C. brenneri* CB5161 is gonochoristic. We obtained both CB5161 and PS1010 from the CGC strain collection and cultured them on OP50 at 20˚C, using methods standard for *C. elegans* (Sulston and Hodgkin 1988).

DNA preparation. Nematode DNA was prepared by two consecutive shearings, first by vortexing and second by needle. For CB5161, 36,864 clones were picked and gridded onto 96 384-well plates; 20-25% of the clones were *C. brenneri* rather than *E. coli* DNA. For PS1010, 100,992 clones were picked and gridded onto 263 384-well plates, and 60-70% of the clones contained *C.* sp. 3 DNA. Both clone libraries had a mean insert size of 36 kb; assuming a genome size of ~100 Mb, like that of *C. elegans* and

C. briggsae (Stein et al. 2003), this gave roughly 3x and 24x genomic coverage for *C. brenneri* and *C.* sp. 3 PS1010. cDNA clones to be used as probes were obtained from: Y. Kohara for the *C. elegans* genes *ceh-13, daf-19, egl-44, egl-46, gcy-8, lin-11, lov-1, nlp-8, osm-5, pkd-2*, and *ref-1*; C. Kenyon for *lin-39* and *mab-5*; W. Wood for *nob-1* and *php-3*; and the Sternberg laboratory for *egl-5*, *egl-30*, and *lin-3*. Probes were radiolabeled by random priming, and fosmids were screened at moderate stringency using otherwise standard methods (Sambrook and Russell 2001).

Sequence analysis. To reconstruct known regulatory motifs, and to see how comparing different numbers of species made motifs more or less detectable, sequences of the *lin-3* anchor cell (ACEL) and *lin-11* uterine enhancer elements (Gupta and Sternberg 2002; Hwang and Sternberg 2004) were linked from *C. elegans* to other species by blocks of identity found with MUSSA. Sequences equivalently positioned around these blocks were then analysed. *lin-11*'s uterine element in *C. elegans*, as defined in WormBase release WS180, is I:10,245,795..10,246,254 (B. Gupta, pers. comm.). Its equivalents were easily found with a large MUSSA block at 22/30 stringency (Figure S8D), and are listed in Table S3. *lin-3*'s ACEL in WS180 is IV:11,059,133..11,059,192 (Hwang and Sternberg 2004); it is invisible to MUSSA at 22/30 stringency, but a 10/10 MUSSA block maps onto one of its two required E-box motifs (Figure S8C), which let us define ACEL equivalents in other species (Table S5).

Nonredundant, statistically overrepresented 6-nt motifs within these regions were generated with YMF (Sinha and Tompa 2002) and Explanators (Blanchette and Sinha 2001). YMF was used to find hexamers, allowing 0 spacers in the middle of a hexamer and a maximum of two degenerate sites within a hexamer. Explanators was then used to find the 5 best nonredundant motifs from a raw YMF output. Both programs were run via Web server (*http://abstract.cs.washington.edu/~saurabh/YMFWeb/YMFInput.pl*) (Sinha and Tompa 2003).

DNA sequence identities were found with *seqcomp* (Brown et al. 2002); we devised the MUSSA software package to adapt *seqcomp* to multiple sequence analysis.

Overrepresented GO terms were identified with the GOstat server (*http://gostat.wehi.edu.au*; Beissbarth and Speed 2004), using a Benjamini and Hochberg correction for multiple testing.

MUSSA (Multiple Species Sequence Analysis). MUSSA will compile on Linux or Mac OS X, given availability of the Fltk graphics library (*http://www.fltk.org*). It has a graphical user interface (GUI) but may also be run at the command line in UNIX-based systems. In the GUI, alignments are visualized

as lines between sequences (red for a direct alignment and blue for a reverse complement alignment), and the sequences are displayed one above another. Using a *seqcomp*-based sliding window algorithm, we varied the threshold of conservation (60-100% identity) and window size (10-30 bp) for identifying conserved regions (Brown 2006; Brown et al. 2002). For the thresholds used in the study, all matches represent a statistically significant enrichment in conservation compared to a random model (Brown 2006). Match threshold and window size, dependent on base pairs, must be integer values; fractional nucleotides are not possible. MUSSA runs all possible pairwise sequence comparisons among two or more (N) genomes, then integrates all pairwise matched features by requiring them to match transitively. Transitivity requires that (for example, in a 3-way comparison with sequence window W and sequences A, B, and C) if W_{AB} and W_{BC} meet the threshold, then W_{AC} must meet the threshold to qualify as a match. Note that individual base pairs are not required to be identical across all pairwise comparisons. Transitivity filtering gives equal weight in the comparison to all participating genomes, and the interactive viewer highlights all relationships that strictly pass the transitivity test. Mussa images were generated by the MUSSA GUI.

MEME. The MEME web interface (*http://meme.sdsc.edu/meme*) was used for submitting short genomic sequences and retrieving overrepresented motifs, with the expectation of zero or one occurrences per sequence.

Transgene design and construction. PCR fusions (Hobert 2002) were generated with Roche Expand Long Template and Expand High Fidelity PCR systems. An additional nested primer, designed to have a T_m closer to those used with the enhancer elements, was used in place of the Hobert nested primer. For the enhancer element side of the fusion, the left primer was reused rather than using a nested primer. The Fire Lab Vector pPD107.94 was used as the template for the Δ*pes-10*::4X-NLS::eGFP::LacZ::*unc-54* sequence.

For mutations of sites, the mutation primers were used with the Stratagene PfuUltra Hotstart on plasmids containing the insert. The mutated and sequenced enhancers were fused to a modified Fire Lab Vector pPD122.53 with YFP replacing the GFP, to give a Δ*pes-10*::4X-NLS::YFP::*unc-54* sequence. Control un-mutated and sequenced enhancers were fused to pPD122.53 with CFP replacing GFP, to give a Δ*pes-10*::4X-NLS::CFP::*unc-54* sequence. The PCR fusion products were used directly for microinjection, and not purified or sequenced following the fusion.

To determine the regions to be reproduced for the expression analysis, the conserved element was buffered by 200 base pairs on either side and additional bases were allowed for enhanced primer picking. Primer3 was used (*http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi*) to select primers, using an optimal T_m of 62˚C and optimal length of 21 bp. BLAST was used to find occurrences of the proposed primers in the genome to screen out popular matches prior to selection in order to prevent non-specific hybridization (*http://www.ensembl.org/Caenorhabditis_elegans/index.html*). The primers termed C and DS are modified from Hobert (2002). Primers, as listed in Table S4, were ordered from Integrated DNA Technologies.

Nomarski imaging. Transgenic animals were viewed with Nomarski optics and a Chroma High Q EnGFP LP, YFP LP, or CFP filter cube on a Zeiss Axioplan, with a 100X oil objective, a 200-watt HBO UV epifluorescence light source, and a Hamamatsu ORCA II digital camera using Improvision Openlab software. ImageJ v1.37 was used to adjust image brightness and contrast and generate overlays. Transgenic lines were fixed in 4% formaldehyde for pre-screening of expression across all stages of life. Live worms on 2% noble agar and 0.1 M sodium azide were then analyzed, described, and imaged.

Confocal imaging. Transgenic animals were fixed with 4% formaldehyde and stained with phalloidin-rhodamine. They were suspended in 2% low-melt agarose and imaged on a Zeiss inverted-410 Axioplan confocal microscope using two excitation lasers (543 nm for the red channel and 488 nm for the green channel) and a 63X oil-dipping objective. Imaging was performed with two monochrome photomultiplier tubes and captured with Zeiss Axiovision software. Brightness and contrast of images were adjusted and multi-channel maximum intensity projections of 0.3 µm spaced sections were created using ImageJ.

Sources of Accession Numbers. *C. elegans* gene accession numbers were taken from WormBase archival release WS180. Vertebrate gene accession numbers, unless otherwise noted, were taken from Ensembl release 47 (Oct 2007).

Supplementary Tables:

Table S1. DNA and predicted protein sequences from *C. brenneri*.

Contig	Contig Length (nt)	Contig Protein	Protein Length (aa)	Predicted Protein

Cbre_JD01	37,836	Cbre_JD01.001	715	WBGene00016652|C44E4.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD01.002	86	WBGene00016655|acbp-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD01.003	422	WBGene00016653|C44E4.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD01.004	4,217	WBGene00016650|C44E4.1 and WBGene00016656|C44E4.7 (2 elegans, 2 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD01.005	640	
		Cbre_JD01.006	920	WBGene00022369|Y92H12BR.3 [*] (1 elegans, 2 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD01.007	177	WBGene00022368|Y92H12BR.2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD01.008	333	WBGene00022371|Y92H12BR.6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
Cbre_JD02	36,856	Cbre_JD02.001	180	
		Cbre_JD02.002	387	
		Cbre_JD02.003	340	WBGene00003977|pes-2 and WBGene00010158|F56G4.3 (2 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD02.004	796	WBGene00016441|C35D10.3 (1 elegans, 3 briggsae, 10 remanei, 6 brenneri).
		Cbre_JD02.005	299	WBGene00016441|C35D10.3 (1 elegans, 3 briggsae, 10 remanei, 6 brenneri).
		Cbre_JD02.006	509	WBGene00016441|C35D10.3 (1 elegans, 3 briggsae, 10 remanei, 6 brenneri).
		Cbre_JD02.007	314	WBGene00016441|C35D10.3 (1 elegans, 3 briggsae, 10 remanei, 6 brenneri).
		Cbre_JD02.008	851	WBGene00016441|C35D10.3 (1 elegans, 3 briggsae, 10 remanei, 6 brenneri).
		Cbre_JD02.009	316	WBGene00016441|C35D10.3 (1 elegans, 3 briggsae, 10 remanei, 6 brenneri).
		Cbre_JD02.010	98	
Cbre_JD03	16,003	Cbre_JD03.001	802	WBGene00008011|C38D9.3, WBGene00008864|F15D4.7, WBGene00012798|Y43F4A.3, WBGene00017185|F07B7.1, WBGene00020724|T23B12.10, and WBGene00021106|W09B7.1 (6 elegans, 4 briggsae, 61 remanei, 1 brenneri).
		Cbre_JD03.002	120	
		Cbre_JD03.003	221	(1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD03.004	46	
Cbre_JD04	20,546	Cbre_JD04.001	403	WBGene00020867|shc-2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD04.002	601	WBGene00020868|T27F7.3 [*] (1 elegans, 1 briggsae,

				1 remanei, 1 brenneri).
		Cbre_JD04.003	121	WBGene00020866|T27F7.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD04.004	127	WBGene00003425|msp-10, WBGene00003432|msp-36, WBGene00003449|msp-56, and WBGene00003463|msp-76 (4 elegans, 3 briggsae, 16 remanei, 1 brenneri).
		Cbre_JD04.005	52	
		Cbre_JD04.006	164	WBGene00004382|rnh-1.0 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD04.007	180	WBGene00004382|rnh-1.0 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD04.008	73	
		Cbre_JD04.009	70	
		Cbre_JD04.010	247	WBGene00007303|rnh-1.3 [*] (1 elegans, 1 briggsae, 1 brenneri).
Cbre_JD05	10,514	Cbre_JD05.001	81	(2 brenneri).
		Cbre_JD05.002	127	(2 brenneri).
		Cbre_JD05.003	1,331	WBGene00021678|Y48G1C.5 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri).
		Cbre_JD05.004	115	WBGene00003097|lys-8 [*] (1 elegans, 1 briggsae, 2 remanei, 2 brenneri).
Cbre_JD06	18,120	Cbre_JD06.001	676	WBGene00020183|T03D3.5 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri).
		Cbre_JD06.002	324	WBGene00017090|E01A2.8 and WBGene00044697|K05F6.11 (2 elegans, 2 briggsae, 1 remanei, 2 brenneri).
		Cbre_JD06.003	231	
		Cbre_JD06.004	284	
Cbre_JD07	66,849	Cbre_JD07.001	1,272	WBGene00000549|cls-2 and WBGene00015580|C07H6.3 (2 elegans, 2 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD07.002	912	WBGene00000537|clk-2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD07.003	513	WBGene00000854|cux-7 [*] (1 elegans, 1 briggsae, 1 brenneri).
		Cbre_JD07.004	105	WBGene00015579|C07H6.2 [*] (1 elegans, 1 briggsae, 1 brenneri, 1 ps1010).
		Cbre_JD07.005	703	WBGene00002986|lig-4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD07.006	95	
		Cbre_JD07.007	252	WBGene00003024|lin-39 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD07.008	78	
		Cbre_JD07.009	42	
		Cbre_JD07.010	68	
		Cbre_JD07.011	54	

		Cbre_JD07.012	202	WBGene00000437|ceh-13 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD07.013	139	(2 brenneri).
		Cbre_JD07.014	141	(2 brenneri).
		Cbre_JD07.015	260	WBGene00022102|Y69F12A.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
Cbre_JD08	27,634	Cbre_JD08.001	341	WBGene00013956|ZK265.3 [*] (1 elegans, 1 remanei, 1 brenneri).
		Cbre_JD08.002	393	WBGene00000639|col-63 [*] (1 elegans, 2 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD08.003	290	WBGene00000433|ceh-8 [*] (1 elegans, 1 remanei, 1 brenneri).
		Cbre_JD08.004	283	WBGene00044094|ZK265.9 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD08.005	189	WBGene00013958|ZK265.6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD08.006	414	WBGene00013957|sre-23 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD08.007	43	
		Cbre_JD08.008	370	WBGene00013959|ZK265.7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
Cbre_JD09	32,968	Cbre_JD09.001	326	WBGene00000603|col-14 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD09.002	1,255	WBGene00011530|T06D8.10, WBGene00016700|C46A5.4, and WBGene00019613|K10B4.1 (3 elegans, 3 briggsae, 4 remanei, 1 brenneri).
		Cbre_JD09.003	479	WBGene00016848|C50F7.10 and WBGene00017103|E02H9.5 (2 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD09.004	417	WBGene00016842|C50F7.1 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD09.005	373	WBGene00011290|R102.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD09.006	149	WBGene00011291|R102.4 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD09.007	184	
		Cbre_JD09.008	266	WBGene00021541|Y42H9B.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD09.009	318	WBGene00016130|C26B2.8 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD09.010	136	WBGene00016129|C26B2.7 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD09.011	335	WBGene00016128|C26B2.6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD09.012	862	WBGene00016124|C26B2.1 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).

Cbre_JD10	46,499	Cbre_JD10.001	49	
		Cbre_JD10.002	472	WBGene00001208|egl-44 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD10.003	508	WBGene00007415|C07E3.4 and WBGene00019020|F57H12.5 (2 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD10.004	78	WBGene00019409|K05F1.8 [*] (1 elegans, 1 briggsae, 1 brenneri).
		Cbre_JD10.005	167	WBGene00000403|casy-1 (1 elegans, 1 brenneri).
		Cbre_JD10.006	822	WBGene00000403|casy-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD10.007	175	
		Cbre_JD10.008	97	
		Cbre_JD10.009	202	
Cbre_JD11	40,423	Cbre_JD11.001	224	WBGene00020424|T10H9.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD11.002	133	WBGene00044779|T10H9.8 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD11.003	418	WBGene00020425|T10H9.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD11.004	108	WBGene00004897|snb-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD11.005	598	WBGene00004062|pmp-5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD11.006	467	(1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD11.007	544	WBGene00017205|F07C4.12, WBGene00017431|F13H6.3, WBGene00019652|K11G9.1, WBGene00019653|K11G9.2, and WBGene00019654|K11G9.3 (5 elegans, 4 briggsae, 6 remanei, 3 brenneri).
		Cbre_JD11.008	574	WBGene00017205|F07C4.12, WBGene00017431|F13H6.3, WBGene00019652|K11G9.1, WBGene00019653|K11G9.2, and WBGene00019654|K11G9.3 (5 elegans, 4 briggsae, 6 remanei, 3 brenneri).
		Cbre_JD11.009	548	WBGene00017205|F07C4.12, WBGene00017431|F13H6.3, WBGene00019652|K11G9.1, WBGene00019653|K11G9.2, and WBGene00019654|K11G9.3 (5 elegans, 4 briggsae, 6 remanei, 3 brenneri).
		Cbre_JD11.010	287	WBGene00001210|egl-46 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD11.011	76	
		Cbre_JD11.012	84	

		Cbre_JD11.013	419	WBGene00019655|K11G9.5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD11.014	75	WBGene00003473|mtl-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD11.015	80	
		Cbre_JD11.016	71	WBGene00020947|W02F12.2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
Cbre_JD12	36,178	Cbre_JD12.001	304	WBGene00001668|gpa-6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD12.002	668	WBGene00009844|cwp-5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD12.003	90	WBGene00003741|nlp-3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD12.004	36	
		Cbre_JD12.005	60	
		Cbre_JD12.006	80	(1 briggsae, 1 remanei, 1 brenneri).
Cbre_JD13	48,934	Cbre_JD13.001	261	WBGene00013891|ZC434.3 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri, 2 ps1010).
		Cbre_JD13.002	203	WBGene00013891|ZC434.3 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri, 2 ps1010).
		Cbre_JD13.003	884	WBGene00002153|irs-2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD13.004	122	(1 briggsae, 1 brenneri).
		Cbre_JD13.005	136	WBGene00007708|C25A1.6 [*] (1 elegans, 1 remanei, 1 brenneri).
		Cbre_JD13.006	316	WBGene00007707|C25A1.5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD13.007	449	WBGene00007706|C25A1.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD13.008	193	WBGene00001442|fkh-10 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD13.009	225	WBGene00007705|C25A1.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD13.010	372	WBGene00006447|tag-72 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD13.011	812	WBGene00002994|lin-5 and WBGene00008508|F01G10.5 (2 elegans, 1 briggsae, 4 remanei, 1 brenneri).
		Cbre_JD13.012	369	WBGene00003000|lin-11 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD13.013	642	WBGene00013860|ZC247.2 and WBGene00013895|ZC434.9 (2 elegans, 2 briggsae, 2 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD13.014	1,747	WBGene00013859|ZC247.1 (1 elegans, 6 briggsae, 18 remanei, 1 brenneri).
Cbre_JD14	34,738	Cbre_JD14.001	139	WBGene00001426|fkb-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).

		Cbre_JD14.002	1,432	WBGene00006490|tag-144 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD14.003	75	WBGene00009496|F36H1.11 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD14.004	117	WBGene00009497|F36H1.12 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD14.005	483	WBGene00002992|lin-3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD14.006	132	WBGene00012382|Y5F2A.1 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD14.007	131	WBGene00012383|Y5F2A.2 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD14.008	78	
		Cbre_JD14.009	450	WBGene00012385|Y5F2A.4 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD14.010	645	WBGene00010882|atgr-7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD14.011	147	WBGene00002344|let-70 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD14.012	541	WBGene00000246|bcc-1 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD14.013	214	WBGene00010883|M7.7 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
Cbre_JD15	13,751	Cbre_JD15.001	204	WBGene00018965|F56D2.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD15.002	420	WBGene00022632|ZC581.2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD15.003	123	WBGene00017299|F09F7.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
Cbre_JD16	19,586	Cbre_JD16.001	152	WBGene00003371|mlc-3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD16.002	1,152	WBGene00016140|rpb-2 and WBGene00017300|F09F7.3 (2 elegans, 2 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD16.003	386	WBGene00017301|F09F7.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD16.004	314	WBGene00017304|F09F7.7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD16.005	87	WBGene00017305|nspb-12 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD16.006	74	
Cbre_JD17	35,362	Cbre_JD17.001	208	(2 brenneri).
		Cbre_JD17.002	318	(2 brenneri).
		Cbre_JD17.003	383	WBGene00008401|D2005.6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD17.004	15	
		Cbre_JD17.005	170	WBGene00003746|nlp-8 [*] (1 elegans, 1 briggsae, 1

				remanei, 1 brenneri).
Cbre_JD18	6,580	Cbre_JD18.001	179	WBGene00007166|B0391.11, WBGene00008014|C38D9.6, WBGene00009836|F47H4.4, WBGene00009837|F47H4.6, WBGene00009838|F47H4.7, WBGene00009840|F47H4.9, WBGene00012566|Y37H2A.6, WBGene00012879|Y45F10C.3, WBGene00015746|C13F10.7, and WBGene00021178|Y9C9A.8 (10 elegans, 4 briggsae, 38 remanei, 4 brenneri).
		Cbre_JD18.002	1,039	WBGene00007166|B0391.11, WBGene00008014|C38D9.6, WBGene00009836|F47H4.4, WBGene00009837|F47H4.6, WBGene00009838|F47H4.7, WBGene00009840|F47H4.9, WBGene00012566|Y37H2A.6, WBGene00012879|Y45F10C.3, WBGene00015746|C13F10.7, and WBGene00021178|Y9C9A.8 (10 elegans, 4 briggsae, 38 remanei, 4 brenneri).
Cbre_JD19	25,578	Cbre_JD19.001	2,149	(1 remanei, 1 brenneri).
		Cbre_JD19.002	443	WBGene00007166|B0391.11, WBGene00008014|C38D9.6, WBGene00009836|F47H4.4, WBGene00009837|F47H4.6, WBGene00009838|F47H4.7, WBGene00009840|F47H4.9, WBGene00012566|Y37H2A.6, WBGene00012879|Y45F10C.3, WBGene00015746|C13F10.7, and WBGene00021178|Y9C9A.8 (10 elegans, 4 briggsae, 38 remanei, 4 brenneri).
		Cbre_JD19.003	1,415	WBGene00004323|rde-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD19.004	528	WBGene00007166|B0391.11, WBGene00008014|C38D9.6, WBGene00009836|F47H4.4, WBGene00009837|F47H4.6, WBGene00009838|F47H4.7, WBGene00009840|F47H4.9, WBGene00012566|Y37H2A.6, WBGene00012879|Y45F10C.3, WBGene00015746|C13F10.7, and WBGene00021178|Y9C9A.8 (10 elegans, 4 briggsae, 38 remanei, 4 brenneri).

Cbre_JD20	38,441	Cbre_JD20.001	468	WBGene00011041|R05H5.7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD20.002	149	WBGene00011038|R05H5.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD20.003	435	WBGene00011039|R05H5.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD20.004	238	WBGene00011040|R05H5.5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD20.005	49	
		Cbre_JD20.006	516	WBGene00011331|T01E8.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD20.007	68	
		Cbre_JD20.008	386	WBGene00004334|ref-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
Cbre_JD21	33,648	Cbre_JD21.001	81	(2 brenneri).
		Cbre_JD21.002	127	(2 brenneri).
		Cbre_JD21.003	1,331	WBGene00021678|Y48G1C.5 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri).
		Cbre_JD21.004	154	WBGene00003097|lys-8 [*] (1 elegans, 1 briggsae, 2 remanei, 2 brenneri).
		Cbre_JD21.005	596	WBGene00020183|T03D3.5 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri).
		Cbre_JD21.006	366	WBGene00017090|E01A2.8 and WBGene00044697|K05F6.11 (2 elegans, 2 briggsae, 1 remanei, 2 brenneri).
		Cbre_JD21.007	371	WBGene00010366|H05L14.1 (1 elegans, 2 briggsae, 3 remanei, 1 brenneri).
		Cbre_JD21.008	381	WBGene00005749|srw-2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD21.009	326	WBGene00008568|F08A8.5 and WBGene00012070|T26H5.8 (2 elegans, 1 briggsae, 2 remanei, 1 brenneri).
Cbre_JD22	33,589	Cbre_JD22.001	427	
		Cbre_JD22.002	73	
		Cbre_JD22.003	156	(7 briggsae, 1 brenneri).
		Cbre_JD22.004	118	(4 remanei, 1 brenneri).
		Cbre_JD22.005	67	
		Cbre_JD22.006	342	(5 briggsae, 1 brenneri).

The names of orthologous *C. elegans* genes, and numbers of orthologous protein-coding genes from other *Caenorhabditis* species, are listed. [*] denotes a strict orthology, as defined in Methods.

Table S2. DNA and predicted protein sequences from *C.* sp. 3 PS1010.

Contig	Contig Length (nt)	Contig Protein	Protein Length (aa)	Predicted Protein

Csp3_JD01	43,544	Csp3_JD01.001	975	WBGene00018721|polh-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD01.002	578	WBGene00004491|rps-22 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD01.003	383	WBGene00017732|F23C8.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD01.004	4,291	WBGene00000396|cdh-4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
Csp3_JD02	87,114	Csp3_JD02.001	931	WBGene00016015|C23G10.8 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD02.002	247	WBGene00004472|rps-3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD02.003	181	WBGene00016011|C23G10.2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD02.004	311	WBGene00004400|rom-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD02.005	98	WBGene00015579|C07H6.2 [*] (1 elegans, 1 briggsae, 1 brenneri, 1 ps1010).
		Csp3_JD02.006	683	WBGene00002986|lig-4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD02.007	210	WBGene00003024|lin-39 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD02.008	368	WBGene00007305|C04G2.2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD02.009	200	WBGene00000437|ceh-13 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD02.010	300	WBGene00021260|Y22D7AR.6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD02.011	621	WBGene00021460|zwl-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD02.012	317	WBGene00021258|Y22D7AR.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD02.013	227	WBGene00021254|Y22D7AL.16 [*] (1 elegans, 1 briggsae, 1 ps1010).
		Csp3_JD02.014	64	WBGene00018363|F42G9.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD02.015	484	WBGene00011407|T04A8.5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD02.016	288	WBGene00011408|T04A8.6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD02.017	1,254	WBGene00011409|T04A8.7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD02.018	485	WBGene00011199|tag-310 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD02.019	131	WBGene00019329|K02F3.6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
Csp3_JD03	47,839	Csp3_JD03.001	1,481	WBGene00006805|unc-73 (1 elegans, 2 briggsae, 2 remanei, 1 ps1010).

		Csp3_JD03.002	491	WBGene00022141|Y71G12B.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD03.003	660	WBGene00016907|C53H9.2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD03.004	355	WBGene00001196|egl-30 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD03.005	181	WBGene00001309|emr-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD03.006	457	WBGene00006461|tag-96 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD03.007	317	WBGene00004743|scm-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD03.008	872	WBGene00022139|tag-305 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD03.009	432	WBGene00001007|dli-1 [*] (1 elegans, 1 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD03.010	361	WBGene00009140|F26A3.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
Csp3_JD04	81,328	Csp3_JD04.001	503	WBGene00000117|alh-11 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD04.002	477	WBGene00001573|gei-16 (1 elegans, 1 ps1010).
		Csp3_JD04.003	949	WBGene00001573|gei-16 (1 elegans, 1 ps1010).
		Csp3_JD04.004	181	
		Csp3_JD04.005	1,332	WBGene00020550|T17H7.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD04.006	167	
		Csp3_JD04.007	177	
		Csp3_JD04.008	191	WBGene00003102|mab-5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD04.009	252	WBGene00015591|C08C3.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD04.010	211	WBGene00001174|egl-5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD04.011	1,086	WBGene00000768|cor-1 and WBGene00007983|C36E8.4 (2 elegans, 2 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD04.012	775	
		Csp3_JD04.013	340	WBGene00003162|mdh-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD04.014	117	WBGene00019509|K07H8.9 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
Csp3_JD05	66,535	Csp3_JD05.001	213	WBGene00004418|rpl-7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD05.002	251	WBGene00018774|F53G12.9 [*] (1 elegans, 1 remanei, 1 ps1010).
		Csp3_JD05.003	1,639	WBGene00003210|mel-28 (1 elegans, 2 briggsae, 2 remanei, 1 ps1010).

		Csp3_JD05.004	1,876	WBGene00002040|hum-7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD05.005	503	WBGene00022709|ZK354.8 [*] (1 elegans, 1 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD05.006	291	WBGene00014083|ZK795.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD05.007	1,195	WBGene00006961|xnp-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD05.008	304	WBGene00012156|ebp-2 [*] (1 elegans, 2 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD05.009	347	WBGene00006447|tag-72 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD05.010	344	WBGene00003000|lin-11 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD05.011	1,077	WBGene00013860|ZC247.2 and WBGene00013895|ZC434.9 (2 elegans, 2 briggsae, 2 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD05.012	335	WBGene00011340|ugt-30, WBGene00015693|ugt-28, and WBGene00021709|ugt-29 (3 elegans, 1 briggsae, 2 remanei, 2 ps1010).
		Csp3_JD05.013	332	WBGene00011340|ugt-30, WBGene00015693|ugt-28, and WBGene00021709|ugt-29 (3 elegans, 1 briggsae, 2 remanei, 2 ps1010).
		Csp3_JD05.014	295	WBGene00013893|ZC434.7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD05.015	82	
		Csp3_JD05.016	1,841	WBGene00000148|aph-2 and WBGene00001337|ers-2 (2 elegans, 2 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD05.017	258	WBGene00013891|ZC434.3 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri, 2 ps1010).
		Csp3_JD05.018	276	WBGene00013891|ZC434.3 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri, 2 ps1010).
		Csp3_JD05.019	271	WBGene00013892|ZC434.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
Csp3_JD06	60,757	Csp3_JD06.001	486	WBGene00005663|srs-2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD06.002	301	WBGene00008147|C47E12.2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD06.003	546	WBGene00008148|C47E12.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD06.004	325	WBGene00022707|ZK354.6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD06.005	441	WBGene00009686|F44D12.9 (1 elegans, 2 briggsae, 3 remanei, 1 ps1010).
		Csp3_JD06.006	528	WBGene00003992|pgl-1 and WBGene00003994|pgl-3 (2 elegans, 1 briggsae, 1 ps1010).
		Csp3_JD06.007	234	WBGene00011746|T13F2.6 [*] (1 elegans, 1

				briggsae, 1 remanei, 1 ps1010).
		Csp3_JD06.008	223	WBGene00002274|lec-11 [*] (1 elegans, 1 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD06.009	83	
		Csp3_JD06.010	510	WBGene00003603|nhr-4 [*] (1 elegans, 1 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD06.011	391	WBGene00002992|lin-3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD06.012	136	WBGene00009497|F36H1.12 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD06.013	1,476	WBGene00006490|tag-144 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD06.014	271	WBGene00001426|fkb-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD06.015	858	WBGene00015571|C07G1.2 [*] (1 elegans, 1 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD06.016	641	WBGene00003838|ocr-1, WBGene00003839|ocr-2, and WBGene00003840|ocr-3 (3 elegans, 3 briggsae, 3 remanei, 1 ps1010).
Csp3_JD07	30,012	Csp3_JD07.001	245	WBGene00015156|B0361.2 [*] (1 elegans, 1 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD07.002	681	WBGene00004905|snf-6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD07.003	351	WBGene00019716|M01G5.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD07.004	27	
		Csp3_JD07.005	138	
		Csp3_JD07.006	849	WBGene00019715|M01G5.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD07.007	340	WBGene00022793|ZK686.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD07.008	218	WBGene00022794|ZK686.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD07.009	657	WBGene00008167|C48B4.1, WBGene00008564|F08A8.1, WBGene00008565|F08A8.2, WBGene00008566|F08A8.3, and WBGene00008567|F08A8.4 (5 elegans, 5 briggsae, 4 remanei, 1 ps1010).

The names of orthologous *C. elegans* genes, and numbers of orthologous protein-coding genes from other *Caenorhabditis* species, are listed. [*] denotes a strict orthology, as defined in Methods.

Table S3. Coordinates of elements in *C. elegans*

A. Coordinates of elements in transgenic assays

Element	5' start with respect to *ceh-13*	3' stop with respect to *ceh-13*	Chromosomal location
N1	-24938	-23974	III:7530646..7531610
N2	-23685	-23080	III:7531899..7532504
N3	-22574	-21944	III:7533010..7533640
N4	-19284	-18587	III:7536300..7536997
N5	-17890	-16593	III:7537694..7538991
N6	-12411	-11977	III:7543173..7543607
N7	-11697	-11106	III:7543887..7544478
N8	-10890	-10195	III:7544694..7545389
N9	-6925	-5805	III:7548659..7549779
N10	-2899	-1784	III:7552685..7553800
N11	-825	-6	III:7554759..7555578
I0	-25687	-24938	III:7529897..7530646
I1	-23974	-23685	III:7531610..7531899
I2	-23080	-22769	III:7532504..7532815
I3	-18587	-17890	III:7536997..7537694
I4	-16593	-12411	III:7538991..7543173
I5	-11977	-11697	III:7543607..7543887
I6	-11106	-10890	III:7544478..7544694
I7	-10195	-6925	III:7545389..7548659
I8	-5805	-2899	III:7549779..7552685
I9	-1783	-826	III:7553801..7554758
W2	-11697	-5805	III:7543887..7549779

B. Coordinates of MUSSA matches in initial study

Element	5' start with respect to *ceh-13*	3' stop with respect to *ceh-13*	Chromosomal location
N1	-24807	-24783	III:7530777..7530801
	-24762	-24735	III:7530822..7530849
	-24677	-24629	III:7530907..7530955
	-24060	-24040	III:7531524..7531544
	-24030	-24006	III:7531554..7531578
N2	-23499	-23450	III:7532085..7532134
	-23365	-23339	III:7532219..7532245
N3	-22460	-22433	III:7533124..7533151
N4	-18832	-18815	III:7536752..7536769
	-18802	-18769	III:7536782..7536815
	-18742	-18719	III:7536842..7536865
N5	-17606	-17578	III:7537978..7538006
N6	-12362	-12338	III:7543222..7543246
N7	-11294	-11251	III:7544290..7544333
N8	-10594	-10561	III:7544990..7545023
	-10541	-10514	III:7545043..7545070
	-10290	-10255	III:7545294..7545329
N9	-6583	-6561	III:7549001..7549023
	-6455	-6433	III:7549129..7549151
N10	-2696	-2669	III:7552888..7552915

	-2572	-2547	III:7553012..7553037
N11	-795	-774	III:7554789..7554810
	-642	-622	III:7554942..7554962

C. Coordinates of MUSSA matches with revised parameters (15-bp window)

Element	5' start with respect to *ceh-13*	3' end with respect to *ceh-13*	Chromosomal location
I0	-25385	-25369	III:7530199..7530215
N1	-24801	-24783	III:7530783..7530801
	-24662	-24632	III:7530922..7530952
	-24060	-24045	III:7531524..7531539
	-24023	-24005	III:7531561..7531579
N2	-23499	-23473	III:7532085..7532111
	-23363	-23342	III:7532221..7532242
N3	-22457	-22433	III:7533127..7533151
N4	-18832	-18815	III:7536752..7536769
	-18799	-18771	III:7536785..7536813
N7	-11288	-11255	III:7544296..7544329
N8	-10290	-10261	III:7545294..7545323
N9	-6583	-6564	III:7549001..7549020
	-6534	-6519	III:7549050..7549065
	-6455	-6437	III:7549129..7549147
N10	-2690	-2675	III:7552894..7552909
	-2569	-2547	III:7553015..7553037
	-1822	-1807	III:7553762..7553777
N11	-795	-778	III:7554789..7554806

D. Coordinates of elements and MUSSA matches in mouse

Element	Type of region	Chromosomal location
MmN3	cloned region	chr6:52115073-52115815
	MUSSA match	chr6:52115286-52115301
MmN7	cloned region	chr6:52143858-52144634
	MUSSA match	chr6:52144162-52144181

(A) These are coordinates for the blocks of sequence used in the transgenic assays that were defined as conserved or not conserved by our initial computational analysis. The conserved regions (N) include the matches defined by MUSSA in the Elegans-group comparisons, given in (B), in addition to flanking sequences. The matches determined by the revised parameters, using a 15-bp window at 100%, are given in (C). Sequence coordinates are in reference to the start of *ceh-13* for the first columns and with respect to Chromosome III for the last column. All coordinates are for WormBase build WS180. The coordinates for the mouse sequences are given in (D). These coordinates are for UCSC July 2007 mouse build.

Table S4. Primer sequences

N1L_fus	**CAAGGCCTGCAGGCATGCAAGC**CCATAACCGAAGCAATTCTCTCA
N1R_XbaI	**ATATCTAGA**TGTTACACCGTGTTCTCCCTCAT
N1L_HinDIII	**TCAAAAAGCTT**CCATAACCGAAGCAATTCTCTCA
N2L_fus	**CAAGGCCTGCAGGCATGCAAGC**TTTTAAGCGTCTGCGTCTGAAGT
N2R_XbaI	**ATATCTAGA**TCTCCACTGAATATCGCCAGTTC
N2L_HinDIII	**TCAAAAAGCTT**TTTTAAGCGTCTGCGTCTGAAGT
N3L_fus	**CAAGGCCTGCAGGCATGCAAGC**GCACCCTAGATCAACAAGCTTCA
N3R_XbaI	**ATATCTAGA**TTTGGCAAAACAATGGTCTCAC
N3L_StuI	**TCAAAAGGCCT**GCACCCTAGATCAACAAGCTTCA
N4L_fus	**CAAGGCCTGCAGGCATGCAAGC**TTAAACGTTTTCTGCCACAAAGG
N4R_StuI	**TCAAAAGGCCT**TTTTGTTCCTAAAAGCGGCAACT
N5L_fus	**CAAGGCCTGCAGGCATGCAAGC**CAAATTCTCAGAGCCACAACACA
N5R_SphI	**GCTGCATGC**TACCCCTGTGCAACTCAACAAAT
N6L_fus	**CAAGGCCTGCAGGCATGCAAGC**AGCCAAATGAAGTGCCAATTTTA
N6R_HinDIII	**TTTACAAGCTT**GCCCATCTTCGAAAATTTTGTTT
N7L_fus	**CAAGGCCTGCAGGCATGCAAGC**TTTTTCTTATTTAACCTGCACCACA
N7L_HinDIII	**TCAAAAAGCTT**GGAATGTCGGAGTCCAAAAGAT
N7R_XbaI	**ATATCTAGA**GGAATGTCGGAGTCCAAAAGAT
N8L_SalI	**CATTAGTCGAC**ACAACTTTCGCCTGTGTCTGTTT
N8R_fus	**CAAGGCCTGCAGGCATGCAAGC**CCCTCTAGACACCTGTTGTTCTTCT
N9L_StuI	**TCAAAAGGCCT**TTTCAAAAGTCGCCTTTACAGTCA
N9R_fus	**CAAGGCCTGCAGGCATGCAAGC**CCCGATTAAAAGTTGTAAGGCAAT
N10L_StuI	**TCAAAAGGCCT**ACTGTAGCCCGACACTGATGTTC
N10R_fus	**CAAGGCCTGCAGGCATGCAAGC**CTATGAGGAGATGGACACGGAGT
N11L_HinDIII	**TCAAAAAGCTT**CTCCTTCTTTTCCCCGTGTCC
N11R_fus	**CAAGGCCTGCAGGCATGCAAGC**AGTGGAGCTCATGCTGGAAAATA
I0L_fus	**CAAGGCCTGCAGGCATGCAAGC**TATGCTGTTCGTTGTCGCTTCT
I0R	TGAGAGAATTGCTTCGGTTATGG
I1L_fus	**CAAGGCCTGCAGGCATGCAAGC**ATGAGGGAGAACACGGTGTAACA
I1R	ACTTCAGACGCAGACGCTTAAAA
I2L_fus	**CAAGGCCTGCAGGCATGCAAGC**GAACTGGCGATATTCAGTGGAGA
I2R	TGAAGCTTGTTGATCTAGGGTGC
I3L_fus	**CAAGGCCTGCAGGCATGCAAGC**AGTTGCCGCTTTTAGGAACAAAA
I3R	TGTGTTGTGGCTCTGAGAATTTG
I4L_fus	**CAAGGCCTGCAGGCATGCAAGC**ATTTGTTGAGTTGCACAGGGGTA
I4R	TAAAATTGGCACTTCATTTGGCT
I5L_fus	**CAAGGCCTGCAGGCATGCAAGC**AAACAAAATTTTCGAAGATGGGC
I5R	TGTGGTGCAGGTTAAATAAGAAAAA

I6L	ATCTTTTGGACTCCGACATTCC
I6R_fus	**CAAGGCCTGCAGGCATGCAAGC**AAACAGACACAGGCGAAAGTTGT
I7L	AGAAGAACAACAGGTGTCTAGAGGG
I7R_fus	**CAAGGCCTGCAGGCATGCAAGC**TGACTGTAAAGGCGACTTTTGAAA
I8L	ATTGCCTTACAACTTTTAATCGGG
I8R_fus	**CAAGGCCTGCAGGCATGCAAGC**GAACATCAGTGTCGGGCTACAGT
I9L	ACTCCGTGTCCATCTCCTCATAG
I9R_fus	**CAAGGCCTGCAGGCATGCAAGC**GGACACGGGGAAAAGAAGGAG
N1mL	TACCGCTGCGGGGA*ACAGTTTCATAAACCTG*AGTTGCTCTGATAGCTGTGATG
N1mR	CATCACAGCTATCAGAGCAACT*CAGGTTTATGAAACTGT*TCCCCGCAGCGGTA
N2-1mL	GAAAGTGAGTGGCGGGG*AGCACAGTTCTGG*AAGATAAATGGGCTCGCGAC
N2-1mR	GTCGCGAGCCCATTTATCTT*CCAGAACTGTGCT*CCCCGCCACTCACTTTC
N2-2mL	GCGTCGCCTTCTTCCTT*TAGTAAAACTGTACTTCGTA*GTGGAGAGAGGGAAAAGAAG
N2-2mR	CTTCTTTTCCCTCTCTCCAC*TACGAAGTACAGTTTTACTA*AAGGAAGAAGGCGACGC
N3mL	GAGACAAACAGCGGGAA*TCAAAGTTCTAATTAAC*CTTCCTCTCACTCTTTCACTCTC
N3mR	GAGAGTGAAAGAGTGAGAGGAAG*GTTAATTAGAACTTTGA*TTCCCGCTGTTTGTCTC
N7mL	AAAAGAGGGTAAAGATT*TCTAAATACCCACGGTAATT*CAACTCTCACCAGACGTACG
N7mR	GTCTGGTGAGAGTTG*AATTACCGTGGGTATTTAGA*AATCTTTACCCTCTTTTCCATC
MmN3L_XbaI	**ACATATCTAGA**TGTTTGCCTCCTGATCTGC
MmN3R_HinDIII	**TCAAAAAGCTT**GAAGTTGATGGCGAAGGAAG
MmN3L_fusion	**CAAGGCCTGCAGGCATGCAAGC**TGTTTGCCTCCTGATCTGC
MmN7L_HinDIII	**TCAAAAAGCTT**GCACTGGAGGAGTCCTAACC
MmN7R_XbaI	**ACATATCTAGA**ACTCCCTTCGACTCCATCTG
MmN7R_fusion	**CAAGGCCTGCAGGCATGCAAGC**ACTCCCTTCGACTCCATCTG
C	GCTTGCATGCCTGCAGGCCTTG
DS	CATTTCCCCGAAAAGTGCCACCTGA
D*	GTGTCAGAGGTTTTCACCGTCAT

##L represents the left primer and ##R represents the right primer. Sequences in bold represent the overlapping region utilized in the fusion or the sequence with a restriction site. Italicized sequences represent mutated regions.

Table S5. Known or predicted coordinates of *lin-3* and *lin-11* genes and their regulatory elements.

Gene/Element	Species	Coordinates

lin-3	*elegans*	IV:11053607..11063483
	briggsae	chrIV:5701665..5708512 [antisense]
	remanei	Supercontig32:284661..291046
	brenneri	CB5161_lin-3.tfa:12411..19047
	sp. 3 PS1010	PS1010_lin-3.tfa:31409..36034 [antisense]
ACEL	*elegans*	IV:11059133..11059192
	briggsae	chrIV:5704301..5704360 [antisense]
	remanei	Contig32.18:21275..21334
	brenneri	CB5161_lin-3.tfa:16249..16308
	sp. 3 PS1010	n/a [5' flank was PS1010_lin-3.tfa:34099..36034; antisense]
lin-11	*elegans*	I:10241073..10255621
	briggsae	chrI:6218293..6230072 [antisense]
	remanei	Supercontig31:626189..635406
	brenneri	CB5161_lin-11.tfa:26842..36289
	sp. 3 PS1010	PS1010_lin-11.tfa:31373..37085
uterine	*elegans*	I:10245795..10246254
	briggsae	chrI:6225822..6226281 [antisense]
	remanei	Contig31.36:12788..13247
	brenneri	CB5161_lin-11.tfa:28812..29271
	sp. 3 PS1010	n/a [5' flank was PS1010_lin-11.tfa:31373..32779]

Sequence data coordinates follow the WS180 release of WormBase or our data; the recent CB3 genome assembly (Hillier 2007) was used for *C. briggsae*.

Table S6. Z-scores of known cis-regulatory motifs in *lin-3* and *lin-11*

Sequence	Site	2-spp	3-spp (+rem)	3-spp (+bre)	4-spp	5-spp
CACCTG	E-box (lin-3)	24.52 [1]	30.04 [1]	30.04 [1]	34.68 [1]	12.23 [1]
ACCCTG	Ftz-F1 (lin-3)	15.72 [2]	19.25 [2]	19.25 [2]	22.23 [2]	8.67 [2]
ATGGGA	LAG-1 (lin-11)	[none]	7.78 [~2]	6.59 [4]	9.28 [2]	8.48 [~2]

Known motifs were analyzed between different species using YMF/Explanators. Z-scores for the motifs represent the number of standard deviations from the mean genomic background frequency, as calculated for nonredundant overrepresented hexamers by YMF/Explanators (Blanchette and Sinha 2001; Sinha and Tompa 2002). The first two motifs were generated from known or predicted *lin-3* ACEL sequences; the third was from the *lin-11* uterine enhancer (Gupta and Sternberg 2002). "2-spp" includes *C. elegans* and *C. briggsae*. "3-spp" includes *C. elegans*, *C. briggsae*, and either *C. remanei* (+rem) or *C. brenneri* (+bre). "4-spp" includes *C. elegans*, *C. briggsae*, *C. remanei*, and *C. brenneri*. "5-spp" includes *C. elegans*, *C. briggsae*, *C. remanei*, *C. brenneri*, and *C.* sp. 3 PS1010.

SUPPLEMENTARY FIGURE LEGENDS

Figure S1: The *C. elegans* Hox cluster

The first two pairs of Hox genes (*ceh-13*/*lin-39* and *mab-5/egl-5*) are transcribed away from each other, leaving a large common 5' region between each pair of genes. The third pair (*php-3/nob-1*) are transcribed in the same direction with little space between the two genes, but possess a large intergenic region 5' of *nob-1*. This third pair has only a single ortholog in the nematode *Pristionchus pacificus*, indicating that this pair may have arisen by duplication (Aboobaker and Blaxter 2003b). The gene order of *ceh-13/lin-39* is flipped with respect to the remaining Hox subclusters on chromosome III, with *lin-39*/Hox5/*Sex combs reduced* more 5' and *ceh-13*/Hox1/*labial* more 3' with respect to the other Hox genes. Large-scale inversions exist even in an intact Hox cluster (e.g., that of *Strongylocentrotus purpuratus*) but might be facilitated in *C. elegans* by the sub-cluster's physical and regulatory isolation (Lemons and McGinnis 2006).

Figure S2: Different MUSSA parameters capture similar but non-identical sets of matches

Changes in window size in 2-way analyses at a constant threshold demonstrate that the (A) 30-bp window appears cleaner than the (B) 20-bp window, which has more crosshatched lines. Changes in window size from a (C) 25-bp window to a (D) 30-bp window at a constant threshold reveal a different set of matches (See also Figure 2E,F). Changes in the included species at a constant threshold (90%) and window size (20 bp) reveal many different matches, as between (B) *C. elegans* and *C. briggsae*; (E) *C. elegans*, *C. briggsae*, and *C. brenneri*; (F) *C. elegans*, *C. briggsae*, and *C. remanei*; (G) *C. elegans*, *C. briggsae*, *C. brenneri*, and *C. remanei*; (H) *C. elegans*, *C. briggsae*, *C. brenneri*, and *C*. sp. 3 PS1010; and (I) *C. elegans*, *C. briggsae*, *C. brenneri*, *C. remanei*, and *C*. sp. 3 PS1010. For the greater number of species, a lower threshold of 85% at the same window size (20 bp) is also shown between (J) *C. elegans*, *C. briggsae*, *C. brenneri*, and *C. remanei*; (K) *C. elegans*, *C. briggsae*, *C. brenneri*, and *C*. sp. 3 PS1010; and (L) *C. elegans*, *C. briggsae*, *C. brenneri*, *C. remanei*, and *C*. sp. 3 PS1010.

Figure S3: Cross-phyla MUSSA and MEME comparisons

(A) 10-way MUSSA analysis of the N7 region between nematodes and vertebrates with a threshold of 15 of 20 bp or 75%. (B) MEME analysis run on the nematode, vertebrate, *B. floridae* (lancelet), *S. mansoni* (trematode), and *H. robusta* (annelid) sequences similar to N3 reveals a number of motifs in common between the sequences. The nematode sequences span 592 bp each and the non-

nematode sequences span 600 bp each. For this figure and for Figures S3C-S3E, the 5 top hits produced by MEME are highlighted, with red, orange, yellow, cyan, and green ordered from best to worst hit. The colors within this image and within Figures S3C-S3E are internally consistent only. (C) MEME analysis run on the nematode and vertebrate sequences similar to N3 reveals a number of motifs in common between the ten sequences. The nematode sequences span 307 bp each and the vertebrate sequences span 600 bp each. (D) MEME analysis run on the nematode and vertebrate sequences similar to N7 reveals only one motif in common between nine of the ten sequences. The remaining motifs are mammal-specific. The nematode sequences span 592 bp each and the vertebrate sequences span 777 bp each, except for frog which spans 827 bp. (E) MEME analysis run on the nematode N3 sequences and *Drosophila* sequences similar to N2-2 (as it is non-orthologous to N3 but conserved between *Drosophila*) reveals a lack of motifs in common between the ten sequences. All the motifs that are present in nematodes are only present in at most half of the *Drosophila*, meaning no motifs were in common throughout. The nematode sequences span 592 bp each and the *Drosophila* sequences span 600 bp each.

Figure S4: The reporter vector drives reproducible background expression

(A) Mouse N7 drives background expression in the intestine (highlighted here with yellow arrows), anterior-most bodywall muscle (green arrows), and head neurons (blue arrows) as seen in MmN7::CFP. The scale bar equals 10 microns. (B) An empty vector drives background expression in the intestine, anterior-most bodywall muscle (yellow arrows), excretory cell, enteric muscle, and anal depressor cell. The scale bar equals 10 microns.

Figure S5: Varying window sizes and species gave different ordering of conservation

Graphs showing the maximum threshold where a match is seen in a MUSSA analysis for a given region. Regions that drove expression are white, while those that did not drive detectable expression are black. (A) Different window sizes result in different maximum thresholds for the different regions in 4-species comparisons (15 bp; 20 bp; 25 bp; 30 bp). (B) Averaging the threshold between different window sizes results in different maximum thresholds for the different regions in 4-species comparisons (15-20 bp; 25-30 bp; 15-20-25-30 bp). (C) Different combinations of species result in different maximum thresholds for the different regions comparisons averaged between 20 and 15 base pair windows (*elegans-briggsae*; *elegans-briggsae-brenneri*; *elegans-briggsae-remanei*; *elegans-briggsae-brenneri-remanei*-PS1010; for *elegans-briggsae-brenneri-remanei* see B). (D) Different combinations of species result in

different maximum thresholds for the different regions comparisons with 15 bp windows (*elegans-briggsae*; *elegans-briggsae-brenneri*; *elegans-briggsae-remanei*; *elegans-briggsae-brenneri-remanei*-PS1010; for *elegans-briggsae-brenneri-remanei* see A). (E) Different window sizes result in different maximum thresholds for the different regions in 4-species comparisons (14 bp; 16 bp; 17 bp; 18 bp; 19 bp; for 15 bp see A).

Figure S6: ROC curves

(A) ROC (receiver operating characteristic; Gribskov and Robinson 1996) curves for variable window sizes in 4-species comparisons (window sizes: 15, 20, 25, 30, 15-20 average) demonstrate that the 15-bp window and 15-20 base pair averaging both give the highest sensitivity for the highest specificity. (B) ROC curves for different window sizes between 20-bp and 14-bp windows, showing that the 15-bp window gives the highest sensitivity for the highest specificity. (C) ROC curves for different combinations of species (15-20 average but variable number of species: *elegans-briggsae*, *elegans-briggsae-remanei*, *elegans-briggsae-brenneri*, *elegans-briggsae-brenneri-remanei*, *elegans-briggsae-brenneri-remanei*-PS1010) demonstrate that a four species comparison gives the highest sensitivity for the highest specificity. (D) ROC curves for different combinations of species (15-bp windows but variable number of species: *elegans-briggsae*, *elegans-briggsae-remanei*, *elegans-briggsae-brenneri*, *elegans-briggsae-brenneri-remanei*, *elegans-briggsae-brenneri-remanei*-PS1010) demonstrate that a four species comparison gives the highest sensitivity for the highest specificity. (E) ROC curves for different averages of window sizes in 4-species comparisons (window sizes: 15-20 average, 25-30 average, 15-20-25-30 average) demonstrate that the 15-20 base pair averaging gives the highest sensitivity for the highest specificity for averaged values.

Figure S7: MUSSA predicts regulatory elements in other genes

MUSSA is capable of identifying cis-regulatory regions in certain other genes when using a 15-bp window with a 100% threshold across 4 species. Shown in red blocks on the top sequence is the region published to drive expression (Okkema et al. 1993); green blocks represent coding regions in (A) *unc-54*, (B) *myo-2*, and (C) *myo-3*.

Figure S8: MUSSA comparisons identify *lin-3* and *lin-11* motifs

(A) Comparison of noncoding *lin-3* gene sequences. Both here and in (B), each gene's boundaries are defined by the nearest 5'- and 3' protein-coding sequences of adjacent genes, encompassing all

flanking DNA (Table S5). The ACEL, a known regulatory motif controlling expression in the anchor cell (Gupta and Sternberg 2002), is marked with a green block; E-box and Ftz-F1 motifs are marked in blue and yellow. Exons (marked in grey) are masked; sequence comparisons are only between non-coding DNA at 22/30 identities/window. Similarities are shown by red or blue lines connecting direct or inverted regions of ungapped identity. Noncoding DNA sequences of the Elegans-group *lin-3* genes are much more similar to one another than to *C.* sp. 3 PS1010 *lin-3*. (B) Comparison of noncoding *lin-11* gene sequences. The uterine element, a known regulatory motif controlling expression in the uterus (Hwang and Sternberg 2004), is marked in green; Su(H)/LAG-1 motifs (Table S6) are marked in blue; other markings are as in (A). For *C. elegans*, a transposon (ZC247.4) was used to define its 5' boundary, which otherwise would extend 9.9 kb further to *csnk-1*. As with *lin-3*, *C.* sp. 3 PS1010 *lin-11* is distinct from others. (C) MUSSA blocks and motifs in and around *lin-3*'s ACEL. Motifs are as in (A). The ACEL lacks large MUSSA blocks but a single 10/10 block links its 3' E-boxes. (D) MUSSA blocks and motifs in and around the *lin-11* uterine element. Su(H) motifs are in blue. Both Su(H)/LAG-1 motifs of *C. elegans* are required *in vivo* (Gupta and Sternberg 2002). A MUSSA block at the 5' fringe of the uterine element links the 5' of the two crucial motifs in four species, with the second Su(H) motif lying outside the block but near it. Another MUSSA block contains a novel motif (in red); it is of unknown significance, but co-occurs with (and is as statistically significant as) Su(H) motifs in this element.

Figure S9: The *ceh-13/lin-39* and *mab-5/egl-5* sub-clusters share a single ungapped sequence alignment

(A) The relative location of the different matches is shown. The match between different Hox clusters is highlighted in red. The autoregulatory sequence identified by Streit et al. (2002) is highlighted in green. The other two MUSSA matches are identified with a 15-bp window and a 20 or 30-bp window and highlighted in yellow and blue, respectively. 164 bp are shown. (B) A MUSSA alignment comparison between *C. elegans* and *C. briggsae ceh-13/lin-39* and *mab-5/egl-5* Hox sub-clusters using a 20-bp window and a 90% threshold. All matches are between the coding sequences, but have been masked here for clarity. At lower thresholds, the matches are entirely noise. (C) By adding additional sequences (the *C. remanei* and *C. brenneri ceh-13/lin-39* sub-clusters and the *C. remanei mab-5/egl-5* sub-cluster), the threshold may be lowered enough to 80% (16/20) that a single real match becomes visible, denoted above the top sequence by an asterisk. The extra lines between sequences are all matches between single and di-

nucleotide repeats. (D) The sequence of this match can be viewed, with each red or blue line denoting a perfectly conserved base. This match overlaps with the first N9 MUSSA match identified in the *ceh-13/lin-39* comparisons.

Figure S10: Genome-wide motif refinements

PWMs, visualized with Weblogo (*http://weblogo.berkeley.edu*) (Crooks et al. 1990), of the N2-1 MUSSA match using the Hox clusters of the 4 species, the two-pass refinement in *C. elegans*, the two-pass refinement in *C. briggsae*, and the two-pass refinement in *C. remanei*.