

HLH-1 AND THE *C. ELEGANS* BODY
WALL MUSCLE TRANSCRIPTIONAL
DIFFERENTIATION NETWORK

Thesis by

Steven Gregory Kuntz

In Partial Fulfillment of the Requirements for the
degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2011

(Defended September 7, 2010)

© 2011

Steven Gregory Kuntz

All Rights Reserved

ACKNOWLEDGEMENTS

My interest in how things work has gone back as far as I can remember. I have always had a desire to take things apart to see not only how they work, but also why they work. I recall being intrigued by the peculiarity of electricity and not willing to accept that electrons move toward a positive charge simply because that is what they do. It was not until graduate school that I was able to really ask these questions again, of why things work the way they do. And I was encouraged to not be satisfied with incomplete answers. I found this very exciting. However, the minutia of everyday research can have a draining effect on the passion for science and discovery. Though learning new things is always thrilling, hoping for an exciting discovery that is all your own is hardly sufficient motivation. So I would like to thank my advisors for keeping me going. Their critiques are necessary for a good education, but their contagious excitement and hopeful outlook were what got me through the days, weeks, months, and years. That excitement – the excitement that a seemingly mundane result may actually matter – is what drives research. So I would like to thank Barbara and Paul for their role in keeping me motivated. There is the joke that a student is not ready to graduate until they are bitter and jaded. I feel that I can graduate happily and hopefully because of the encouragement of my professors.

Also necessary is varied scientific criticism. I would like to thank Ellen Rothenberg, for insightful and rigorous biological inquiries; Angela Stathopoulos, for her original and detailed perspective; and Michael Elowitz for his interest and curiosity. I would especially like to thank Ali Mortazavi for his computational help and critical feedback, Andrea Choe for her experimental assistance and reliability, Emily Riggall for her experimental dedication, Ryan Baugh for his scientific suggestions, Mihoko Kato for her help with

writing, Chris Hart for his experimental insight, Gilberto Hernandez for teaching me persistence, Jennifer Green for recommendations for my talks, Leslie Dunipace for demonstrating efficiency, Tristan De Buyscher for his Mussa algorithm, Erich Schwarz for his command of the literature, Brian Williams for his technical expertise, John DeModena for his experimental creativity, Diane Trout, Brandon King, and Henry Amrhein for their computational help, Igor Antonshechkin for his bioinformatics help, Lorian Shaeffer for making libraries, and Chiraj Dalal for being an excellent sounding board.

My family, who have always been supportive, were instrumental in guiding me toward research. I would like to thank my sister for encouraging the aesthetic and artistic side of me, my father for his instilling in me that interest in how things work, and my mother for teaching me that simply because an answer is convenient does not make it correct. I would like to thank my wife Binita for her help and understanding throughout my seven years here. Without her support and encouragement I would have had a much more difficult time making it through.

– Steven Gregory Kuntz

ABSTRACT

To understand the structure and function of gene regulatory networks, it is important to first catalogue the components. Measurable constituents of networks include cis-regulatory elements, identified by their conservation and ability to drive expression; transcription factor binding motifs, identified by protein binding; transcription factors, identified by their necessity in network function; and target genes, identified by their conditional expression. The heart of a regulatory network is the transcription factor, which is dedicated to its role in the network. Transcription factors must be activated and regulate downstream targets in a discrete and reproducible fashion. Any deviation in network function may result in the collapse of the network and death of the animal. Thus, a network must be robust enough to function under a variety of biological conditions. However, network redundancies are inefficient in terms of fitness and lost during the course of evolution. The network structure and function reflects these evolutionary realities: strong sequence conservation of cis-regulatory elements coupled with widespread stochastic transcription factor binding, and ancient transcription factor conservation coupled with overlapping activation of targets. The evolution of functional transcription factor networks therefore must be a balance between conservation and flexibility.

TABLE OF CONTENTS

Acknowledgements	iii
Abstract.....	v
Table of Contents	vi
Chapter I: Toward the understanding of a transcription factor's function in the context of the <i>C. elegans</i> embryonic body wall muscle differentiation network	1
Introduction	1
Network Components	2
Network Structure	18
Figures	24
References	27
Chapter II: Muligenome DNA sequence conservation identifies Hox cis-regulatory elements.....	1
Abstract.....	1
Introduction	1
Results	3
Discussion	8
Methods	11
References	12
Chapter III: Synthetic PAT screen reveals additional myogenic transcription factors	1
Abstract.....	1
Introduction	2
Results	6
Discussion	14
Materials and Methods.....	17
Tables	19
Figures	24
References	34
Chapter IV: Genome-wide studies of HLH-1 regulation and binding in <i>C. elegans</i> myogenesis reveal regulatory interaction between the body wall muscle and non-striated muscle differentiation networks	1
Abstract.....	1
Introduction	2
Results	7
Discussion	22
Materials and Methods.....	29
Tables	32
Figures	38
References	47

Appendix A: Supplement to Kuntz et al. (2008) “Multigenome DNA sequence conservation identifies Hox cis-regulatory elements,” <i>Genome Research</i> 18(12):1955-68.	1
Appendix B: Broitman-Maduro et al. (2009) “The NK-2 class homeodomain factor CEH-51 and the T-box factor TBX-35 have overlapping function in <i>C. elegans</i> mesoderm development,” <i>Development</i> 136(16):2735-46.	1

Chapter 1: Introduction

Toward the understanding of a transcription factor's function in the context of the *C. elegans* embryonic body wall muscle differentiation network

To understand the structure and function of gene regulatory networks, it is important to first catalogue the components. Measurable constituents of networks include cis-regulatory elements, identified by their conservation and ability to drive expression; transcription factor binding motifs, identified by protein binding; transcription factors, identified by their necessity in network function; and target genes, identified by their conditional expression. The heart of a regulatory network is the transcription factor, which is dedicated to its role in the network. Transcription factors must be activated and regulate downstream targets in a discrete and reproducible fashion. Any deviation in network function may result in the collapse of the network and death of the animal. Thus, a network must be robust enough to function under a variety of biological conditions. However, network redundancies are inefficient in terms of fitness and lost during the course of evolution. The network structure and function reflects these evolutionary realities: strong sequence conservation of cis-regulatory elements coupled with widespread stochastic transcription factor binding, and ancient transcription factor conservation coupled with overlapping activation of targets. The evolution of functional transcription factor networks therefore must be a balance between conservation and flexibility.

Cis-regulatory elements

The easiest, and generally cheapest, approach to identify components of a transcriptional regulatory network is informatic. Bioinformatic techniques are very useful when precisely applied for their intended purpose. One primary utility of bioinformatic tools is the identification of experimental targets, such as transcription factor binding motifs and cis-regulatory elements, that function in gene regulatory networks. Though these motifs and cis-regulatory elements are intimately related, they are not functionally identical and require very different techniques for their identification.

Motifs are generally the whole or part of a transcription factor's binding site. They tend to be very short, generally 6-20 base pairs (Sandelin et al., 2004), and occur rather frequently throughout a genome by statistical chance alone. As such, the informatic techniques for identifying such elements require a paring down of the investigated sequence. Sequence reductions are generated from genomic data, such as expression data from microarray experiments (Fox et al., 2005; Fox et al., 2007; Gaudet and Mango, 2002; Guhathakurta et al., 2002; GuhaThakurta et al., 2004). The sequences are then further limited, often by selecting only the first roughly kilobase 5' of the gene. This way the promoters from all genes known to be co-expressed can be compared under the assumption that they are also in part co-regulated. Commonalities are identified and statistically prevalent sequences can be identified as binding motifs (Bailey and Elkan, 1994; Hertz and Stormo, 1999; Pavesi et al., 2004). With this technique a number of important short regulatory sequences in *C. elegans* have been identified, including sequence motifs important in the developmental regulation of muscle, neurons, the gut, and the pharynx (Ao et al., 2004; Etchberger et al., 2007; Gaudet et al., 2004; McGhee et al., 2007; Pauli et al., 2006; Wenick and Hobert, 2004; Zhao et al., 2007a). Algorithms

designed for such analyses must discard large portions of the sequence that lack identifiable common motifs to avoid sequences controlling unrelated regulation.

Cis-regulatory elements are related to motifs but notably different, despite occasional ambiguities in the non-standardized vocabulary. They are typically longer than a single protein-binding site, belying their more complex character. Most transcription factors bind to the genome in concert with other transcription factors and need to recruit various other proteins, such as histone deacetylases, histone acetyltransferases, RNA Polymerase II, and other transcription factors. As such, it is reasonable to assume that the DNA binding is likewise more complex than can be captured with a single motif. Many cis-regulatory elements will even continue to function if some of their component motifs are mutated (Kuntz et al., 2008). With multiple proteins binding to a cis-regulatory element and some proteins having very flexible binding sequence preferences, the extended sequence surrounding a motif will vary from instance to instance. This property will make genome-wide identification of cis-regulatory elements based on sequence comparisons between co-expressed genes virtually impossible. Thus, a different approach is needed.

Cis-regulatory elements are an aggregation of motifs and supporting sequences that are frequently functionally conserved (Brown et al., 2002; Cameron et al., 2005). Sequence conservation analysis uses evolutionary principles to identify biologically important DNA sequences (Brown et al., 2007). Within a clade of developmentally and physiologically similar organisms, cis-regulatory elements controlling a shared network should be preferentially conserved (Tagle et al., 1988). Once a functional sequence has formed, there is evolutionary pressure to maintain the sequence and its binding-dependent

function while the surrounding ‘non-binding’ sequence may mutate more freely (Brown et al., 2002; Cameron et al., 2005). Taking advantage of this conservation, sequences can be compared between species to identify what is preferentially conserved and what more freely mutates. Successfully finding a functional element conserved between species depends on the character and frequency of mutations within the cis-regulatory element and in flanking sequences. There is no need for the entirety of the cis-regulatory element to be highly conserved, meaning that a conserved sequence may cover just a portion of the regulatory region. Thus, not every transcription factor-binding motif is conserved within the regulatory element; rather, some have a tendency to move around as they are replicated and copies are lost during evolution (Hare et al., 2008).

There are a number of computational techniques that can highlight these conserved regions. Some algorithms, such as BLAST, allow for variations in spacing in order to align the different sequences (Korf et al., 2003). This is very useful in certain situations, but given strict steric and structural requirements for most protein-DNA interactions, cis-regulatory elements have few spacing variations. As such, finding regions with preferential selection against insertions and deletions can capture cis-regulatory elements (Brown et al., 2005). Among the simplest conservation algorithms is the sliding window comparison. Programs like Family Relations and Mussa (Brown et al., 2002; Brown et al., 2005; Kuntz et al., 2008) will compare sequences to identify where both base conservation and spacing have been maintained.

All conservation based comparative algorithms are very dependent on the selection of species used for the comparisons. Species that are too distant may have very different methods of regulating a gene, and therefore may share no cis-regulatory

elements. Species that are too close together may not have significant enough divergence between regulatory regions, complicating the separation of functional and non-functional sequence. As more genomes become available with modern sequencing and mapping techniques, the utility of such techniques has become more apparent across different phyla (Cliften et al., 2003; Kato and Sternberg, 2009; Kellis et al., 2003; Krek et al., 2005; Kuntz et al., 2008; Pennacchio et al., 2006; Xie et al., 2005; Yuh et al., 2002). Comparing orthologous cis-regulatory DNA sequences from three or more roughly equidistant genomes has advantages over comparing only two genomes because each additional genome increases evolutionary divergence and thus total mutational distance (Boffelli et al., 2004; Eddy, 2005; Sinha et al., 2004; Stone et al., 2005). The third or fourth genome sequence lowers the frequency of false positive regions, allowing small but functional cis-regulatory sequences to be detected (Kuntz et al., 2008).

Though fewer nematodes have been sequenced than flies or vertebrates, they have so far proven very conducive to such comparative analyses (Kato and Sternberg, 2009; Kirouac and Sternberg, 2003; Kuntz et al., 2008; Ririe et al., 2008). One advantage of *C. elegans* is the relative compactness of the genome, just over 100 million bases, meaning that most intergenic sequences are shorter and more easily compared. However, the genome is not too compact, allowing for spacing between cis-regulatory elements that is lacking in *Drosophila* (Peterson et al., 2009). Within the nematodes, the advent of genomic sequences for *C. remanei*, and *C. brenneri* have made such comparisons possible that were extremely difficult when only *C. elegans* and *C. briggsae* were available (Kirouac and Sternberg, 2003). Additional species should prove useful in order to cover conservation across a wider array of regulatory elements. Distantly related

species such as *C. sp. 3* PS1010 are useful solely for very highly conserved regulatory elements (Kuntz et al., 2008). Comparisons within vertebrate groups have identified a large number of conserved, non-coding regions of DNA with unknown function (Ahituv et al., 2007; Bejerano et al., 2004; Boffelli et al., 2004; Ovcharenko et al., 2005). Without knowing the time and location of functionality, these elements are difficult to test in vertebrates, contrasting with the relative simplicity of testing them in *C. elegans* and observing across all life stages and tissues.

The utility of any bioinformatic technique must be tested experimentally. With motifs, testing must verify both the ability for a transcription factor to bind (when the corresponding factor is known) and for the motif to drive or silence expression. Similarly, the ability of a cis-regulatory element to drive expression may be tested with *in vivo* reporters. Such positive assays are rarely coupled with negative controls to test their predictive efficacy. By testing both regions predicted by conservation to possess regulatory elements and regions predicted to be devoid of such elements we were able to estimate the efficiency of a sliding window sequence conservation algorithm (Kuntz et al., 2008). The Hox cluster proved to be an ideal target for such analysis, due in part to its relatively high levels of conservation and regulatory complexity. With its large introns and a bidirectional promoter, the Hox cluster has a number of ambiguities regarding its regulation.

Hox gene clusters are present throughout bilateria and are crucial for patterning and development. Their function and regulation is partially conserved across phyla (Frasch et al., 1995; Garcia-Fernandez, 2005; Haerry and Gehring, 1997; Malicki et al., 1992; Popperl et al., 1995; Streit et al., 2002), though identifying cis-regulatory elements

with function conserved across phyla has been rare (Haerry and Gehring, 1997; Kuntz et al., 2008; Streit et al., 2002). Their regulatory elements are likely intercalated, both keeping the cluster together and complicating regulatory dissections (Kuntz et al., 2008; Olson et al., 1996). Due to multiple rounds of genome duplication, vertebrates have multiple Hox clusters, with mammals having four independent clusters of nine to eleven genes each (Lemons and McGinnis, 2006). Insects have only a single cluster, but it consists of twelve genes due to internal duplications (Garcia-Fernandez, 2005). Nematodes have a single cluster with only six Hox genes. Many genes and millions of nucleotides divide the *C. elegans* Hox cluster into three sub-clusters of two genes each: *ceh-13* and *lin-39*, *mab-5* and *egl-5*, and *nob-1* and *php-3* (Figure 1) (Aboobaker and Blaxter, 2003; Lemons and McGinnis, 2006). Hox gene expression in *C. elegans* is very complicated and the corresponding regulation in *C. elegans* retains the complexity of all bilaterian Hox clusters, just with fewer genes (Aboobaker and Blaxter, 2003; Clark et al., 1993; Kuntz et al., 2008; Maloof and Kenyon, 1998; McKay et al., 2003; Stoyanov et al., 2003; Streit et al., 2002; Wagmaister et al., 2006; Wang et al., 1993). The *lin-39/ceh-13* sub-cluster is still large by *C. elegans* standards, making it a biologically interesting locus yet tractable for cis-regulatory dissection.

The regulation of *lin-39/sex combs reduced/Hox5* and *ceh-13/labial/Hox1* has been studied by a number of groups. Several studies dissected the introns and 5' promoter of either genes, but only looked up to eight kilobases upstream of the start site (Stoyanov et al., 2003; Streit et al., 2002; Wagmaister et al., 2006). Through the use of comparative sequence analysis, highly conserved sequences were identified even in the center of the locus, far from either gene and missed by previous dissections (Kuntz et al., 2008). By

modelling the dissection around conserved regions, we were able to identify the majority of regulatory elements in the cluster (77%) with a high degree of efficiency, including those at the center of the sub-cluster. All of the conserved regions drove expression and only three less-conserved regions drove expression. This efficiency is on the order of more successful mammalian regulatory element predictions (Prabhakar et al., 2006).

For highly expressed and well-conserved target genes, such as the muscle myosins, conservation is important and captures the cis-regulatory regions in *unc-54*, *myo-3*, and *myo-2* (Okkema et al., 1993) just as well as in the Hox cluster. However, regulatory elements can sometimes escape discovery. Elements controlling behavioral genes will likely change much more rapidly and very closely related species or even strains might be necessary to identify the preferentially conserved sequences. Nevertheless, for very well conserved gene functions, cis-regulatory elements are extremely well conserved, even across phyla (Kuntz et al., 2008; Pennacchio et al., 2006). With the proper complement of sequenced species or strains most regulatory elements should be discoverable.

Transcription factor binding sites

The question of cis-regulation and binding motifs may also be approached from the other side, via the transcription factors. This approach may be taken both *in vivo* and *in vitro*. *In vitro* techniques such as yeast one-hybrid assays allow screening of diverse samples of sequence to identify where a particular transcription factor may bind (Shim et al., 1995). This has proven quite useful with proteins that have strong binding to specific

sequences, such as the bHLH proteins. In *C. elegans*, yeast one-hybrid analyses have identified the different E-boxes that each bHLH protein will bind (Grove et al., 2009), giving a strong starting point for both informatic and experimental genome-wide analyses of transcriptional regulatory networks. Most of the bHLH binding sites are surprisingly similar. Proteins that specify and regulate such different tissues as neurons and muscle, CND-1 and HLH-1 respectively, bind to essentially the same sequence motif, CAGCTG (Grove et al., 2009). This contrasts with the bHLH Twist family factor HLH-8, which has a distinctly different target sequence, CATATG. This useful information helps to inform *in vivo* experimentation.

To investigate binding *in vivo*, chromatin immunoprecipitation (ChIP) allows extraction of a transcription factor along with the DNA to which it is bound. It has been used both on a small scale, looking at gene expression at specific loci (Lei et al., 2009; Oh et al., 2006), and on a genome-wide scale (Whittle et al., 2008; Zhong et al., 2010). With modern microarray and high-throughput sequencing technologies, it is possible to use ChIP to extract transcription factor bound DNA and either hybridize it to a microarray (ChIP-chip) (Whittle et al., 2008) or sequence it (ChIP-sage and ChIP-seq) (Zhong et al., 2010). The results in *C. elegans* differ from those in vertebrates, likely due to two main factors. First, the nematode sequences are very compact – roughly 30 times smaller than the human genome while retaining roughly the same number of genes – so the sequencing read density is much higher when using the same techniques. Consequently binding sites are very close together and may at times be difficult to distinguish. Secondly, the background sequence of *C. elegans* is extremely AT-rich, especially compared to many enhancers, promoters, and coding sequences which are

relatively AT/GC-normal. This leads to a significant sequencing bias, as very GC-poor sequences tend to not be sequenced as deeply with the modern high-throughput sequencing systems.

Through the use of ChIP-seq to analyze genome-wide binding, all cis-regulatory elements targeted by a single transcription factor may be sampled. Antibody limitations – which can be in part counteracted through the use of transgenic tags (Zhong et al., 2010) – and scaling difficulties complicate analyses, but ChIP-seq serves as an important counterpart to purely computational techniques.

Experimental findings have reflected computational predictions. Very short transcription factor binding motifs will by statistical chance appear fairly frequently within the genome. Confounding factors, such as chromatin density and the binding of accessory proteins, may temper protein binding to motifs. Nonetheless, with many motifs throughout the genome it has been predicted that proteins would bind in more places than where they would be useful. The ChIP-seq results have supported this prediction (Cao et al., 2010; Zhong et al., 2010). Numerous binding sites are near genes that have little to nothing to do with the transcription factor's regulatory targets. Observations in *Drosophila* have shown that there is a significant level of binding in both active and inactive regions of the genome (Li et al., 2008). Such superfluous binding is widespread, but may prove to be fickle as non-functional sites appear and disappear. Whether the binding at these non-regulatory sites is maintained may be determined in the future by studying evolutionary conservation of regulatory and non-regulatory binding sites.

Transcription factors

Identification of where transcription factors bind can give a strong understanding of components involved in gene activation. However, no transcription factor acts in isolation; they perform their functions in concert with numerous other transcription factors. This cooperation can help share the duties or add nuance to the function under different conditions. It can also make studying them more difficult. In some networks, knocking out a single protein can completely shut down a network. Common examples of this include genes involved in fate specification, such as *pal-1*, which is necessary to determine the fate of the C, D, and MS lineages in the embryonic worm (Baugh et al., 2005a). Likewise, *pha-4* is necessary for pharyngeal development and directly activates many of the pharyngeal genes (Gaudet and Mango, 2002; Mango et al., 1994). Without *elt-1* epidermal tissue specifies as mesodermal tissue (Spieth et al., 1991). In other systems the network does not shut down. In these cases regulatory factors can be knocked out, with anything from a minor to a lethal effect, without halting fate specification or tissue differentiation. The body wall muscle differentiation network is a prime example.

There are multiple types of muscle in the nematode, each with their own transcription factors (Figure 2), including pharyngeal muscle, non-striated muscle (including enteric and sex specific muscles), and the body wall muscles, which are analogous to the skeletal muscle of vertebrates (Chen, 1994 #17, Fukushige, 2006 #18). Like skeletal muscle, the nematode body wall muscle is responsible for locomotion and is the most prevalent muscle tissue in the animal, making the development of body wall muscle relatively easy to monitor. Like many transcriptional terminal differentiation networks, the muscle differentiation network may be a reinforced feed-forward system

(Davidson 2006, Fukushige, 2006 #18). Without inhibitory feedback, such systems will not cease functioning once initiated despite detrimental mutations. Knocking out any one factor will not halt differentiation (Chen et al., 1992; Fukushige et al., 2006; Harfe et al., 1998a). This is analogous to vertebrates where a complement of *hlf-1*/MRF orthologs (MyoD, myogenin, MRF4, and Myf-5) may all be individually knocked out without muscle formation halting (Braun et al., 1994; Kassar-Duchossoy et al., 2004; Rawls et al., 1998; Rudnicki et al., 1992; Wang and Jaenisch, 1997; Zhang et al., 1995). Since these genes are orthologs, one gene may take the place of another. If all four are missing, muscle differentiation cannot continue (Valdez et al., 2000). However, in *C. elegans* *hlf-1* is the only copy of the MRF family of genes, making its lack of necessity in differentiation both intriguing and experimentally tractable.

To study a transcriptional network whose function is buffered against mutation, it is necessary to first determine what factors are crucial for proper muscle development. Knocking out multiple transcription factors can reveal interactions, as the phenotype is only visible in the presence of other defects. This can be seen in epidermal patterning with a synthetic multi-vulva phenotype (Lu and Horvitz, 1998). Here a single mutation gives no phenotype because a secondary pathway properly specifies patterning. When the secondary pathway is compromised by a second mutation, the phenotype arises, in this case a secondary vulva. Similar analysis can be performed in muscle. The paralysis at the two-fold stage, or PAT, phenotype indicates that no muscle has formed in the worm, but that development has up to that point proceeded successfully (Waterston, 1989). This is because muscle is necessary for body elongation and without it the animal remains short, compact, and horseshoe-shaped. Only a small percentage of mutant animals will exhibit a

PAT phenotype. *hll-1(cc561)* mutants exhibit this phenotype less than 3% of the time. Though the muscle in most of the animals is morphologically defective, it still differentiates and twitches (Chen et al., 1994). By knocking out a second transcription factor like *unc-120*, the majority of animals will exhibit the PAT phenotype (Baugh et al., 2005b; Fukushige et al., 2006). Any gene necessary to complement the *hll-1* differentiation pathway will give a PAT phenotype if knocked out in conjunction with *hll-1*.

With this analysis conducted in *C. elegans* only one MRF (not four) needs to be knocked out in the screening background and RNAi can knock down genes without the need for crosses or knock-outs. Therefore, RNAi screens are ideal to expand the repertoire of myogenic factors. Several myogenic factors have been identified this way and have provided insight into muscle development. *hnd-1* and *ceh-51* were both identified this way, with the bHLH protein *hnd-1* playing a role in C and D lineage muscle specification (Baugh et al., 2005b) and the NK-2 class homeodomain factor *ceh-51* we identified controlling early muscle specification in the MS lineage (Broitman-Maduro et al., 2009). Both of these factors interact strongly with both *hll-1* and *unc-120* but do not interact strongly with each other, reflecting a lineage-dependent subdivision within muscle specification.

We identified other transcription factors with these synthetic PAT phenotype screens, including factors involved in fate decisions and general network regulation, such as *ceh-20*, *ceh-49*, *hmg-1.2*, *grh-1*, and *lin-1*; factors involved in general transcriptional machinery, such as *tbp-1*; and factors involved in other networks whose role in muscle differentiation is unclear, such as *cnd-1*, *sex-1*, and *sdc-1*. Because the transcription

factors regulating muscle differentiation are conserved across phyla (Fukushige et al., 2006), it is possible that our targets will serve as important myogenic factors in either vertebrates or insects as well.

Expression targets

The network counterpart to the transcription factor cohort is the collection of expression targets. These are the genes that are activated or repressed by the network. Several techniques can identify these targets in the tissues the network controls. To do this, nematodes can be studied as an intact organism without the need for cell cultures or immortalized cell lines. Depending on the specific measurement being done, whole animal cell heterogeneity can be a modest technical problem with minor effects on sensitivity, or it can completely confound useful interpretation. The difficulty with an intact animal is obtaining pure samples that are not contaminated with other cell types. Several different approaches have been successful, though each with caveats. Approaches include tagging all mRNA from the desired tissue, physically isolating only the desired tissue, forcing all cells to convert to the tissue via over-expression of a specification factor, and preventing other lineages from forming by knocking out other specification factors.

Tagging mRNA requires a transgenic animal with a tissue-specific protein. The tag binds to mRNA and can be extracted (Roy et al., 2002). Similarly, a fluorescent transgenic tag may be added to the particular tissue, which will then allow embryonic cells to be dissociated and sorted, keeping the desired cell type for culture and analysis

(Fox et al., 2005; Fox et al., 2007). This procedure is the nearest approximation in *C. elegans* to cell culture. As the majority of material is discarded, this technique is ideal for experimentation where small amounts of material are needed. It is also limited to embryonic stages as dissociation of larval or adult cells is very difficult.

Other techniques involve turning all the tissue of the animal into the desired tissue. Over-expression of HLH-1 can induce non-muscle cells to differentiate as muscle (Fukushige and Krause, 2005), though this can have binding and experimental consequences (Fox et al., 2008). Overexpression of any of the myogenic factors can induce at least some cells to become muscle. This can be seen in the overexpression of HLH-1, UNC-120, HND-1, HLH-8, FOZI-1, and CEH-51 (Amin et al., 2007; Broitman-Maduro et al., 2009; Fukushige et al., 2006; Harfe et al., 1998b). HLH-1 is by far the most potent and the only factor that is exclusively expressed in all the body wall muscle (Figure 2) (Fukushige and Krause, 2005). HLH-1 overexpression in otherwise wild type animals leads to what appears to be the fate transformation of all other cells to body wall muscle. UNC-120 and HND-1 both require several permissive mutations to allow fate transformations, presumably to override checks (Fukushige et al., 2006). Since UNC-120 is expressed in non-striated muscle (Hunt-Newbury et al., 2007) and HND-1 is expressed in both the somatic gonad and the germline (Mathies et al., 2003), it is possible that they drive the expression of those cell types as well. CEH-51 drives fate transformations to both body wall muscle and pharyngeal muscle when overexpressed in the AB lineage, where it is not normally expressed (Broitman-Maduro et al., 2009). HLH-8 overexpression in the embryo leads to some cells expressing sex specific muscle genes (Harfe et al., 1998b; Wang et al., 2006; Zhao et al., 2007b). FOZI-1, which acts

redundantly with HLH-1 in post-embryonic body wall muscle, can also drive expression of body wall muscle targets when expressed in non-native tissues (Amin et al., 2007).

Any overexpression approach could increase the number of false-positive binding sites when looking directly at HLH-1 binding. Additionally, such an approach would not be useful when looking at mutations that affect the muscle enhancement phenotype, such as with *hh-1* mutants. Another question that may arise is what exactly comprises muscle in the overexpression mutants. Though the cells do form muscle proteins, it is not clear that cells that normally would produce another tissue do not still express at least a subset of other terminal target proteins.

A final approach is to reduce the expression of unwanted specification factors, thus permitting normally repressed factors to take over the specification. This permissive process arguably allows modification of much of the animal without a need for sorting material and without interference of overexpressed transcription factors. Without the primary specification factors, it is unlikely that downstream targets native to an undesired tissue will be expressed. Rather, by modifying the lineages the cells should be identical to the desired targets. This uniquely nematode technique is dependent on a predetermined lineage.

The problems of this approach include the limitation in available tissue, interference from unknown specification factors, and the need to knock down or knock out multiple genes simultaneously. Most lineages or tissues require multiple gene knock down to isolate them. For instance, knocking down *mex-3*, the repressor of *pal-1*, prevents the AB lineage from forming (Figure 3). However, this does not fully wipe out

any tissue type and leaves the EMS, C, and D lineages (Draper et al., 1996; Hunter and Kenyon, 1996). Instead, the former AB lineage now mimics the C and D lineages. This effectively increases both the amount of muscle and epidermis in the animal. Some tissues can be wiped out relatively easily. Knocking down the GATA factor *elt-1* gets rid of all epidermis (Michaux et al., 2001). What would have become epidermis in the C-lineage now becomes muscle. Conversely, muscle can be easily removed. *pal-1* can be knocked down to get rid of the EMS, C, and D lineages, thus getting rid of all muscle save one cell (Hunter and Kenyon, 1996). These transformations have proven useful in small-scale studies through the use of mutations and balancers (Baugh et al., 2005a; Baugh and Hunter, 2006; Broitman-Maduro et al., 2009). Due to the lethality, it is difficult to scale the lineage control up unless RNAi is used. Luckily, the RNAi against these factors is very effective (Baugh et al., 2005a). Simply by knocking down one gene, *mex-3*, the amount of muscle can be more than doubled. By knocking down *mex-3*, *skn-1*, and *elt-1*, all but the germline can be transformed into muscle (Baugh et al., 2005a; Blackwell et al., 1994; Bowerman et al., 1992; Draper et al., 1996; Michaux et al., 2001). The downside to such a transformation is two-fold. Simultaneously knocking down three genes on a large scale is typically difficult (Gonczy et al., 2000). This can be addressed in part by concatenating the different RNAi transcripts. Secondly, all of the new muscle mimics the C and D lineages rather than the MS lineage, which usually contributes one third of the adult worm muscle (Sulston et al., 1983). This is caused by the fact that *skn-1* is necessary for the EMS lineage to develop, which gives rise to epidermis, muscle, and the intestine (Blackwell et al., 1994; Bowerman et al., 1992). This bias must be taken into account.

These various techniques can generate a muscle-rich animal or isolate exclusively muscle from the animal. Once the isolation is complete, the transcriptome may be catalogued, either by taking the absolute measure of what is present in the muscle (or other isolated tissue) (Fox et al., 2005; Fox et al., 2007) or by selecting what is enriched over whole animals (Baugh et al., 2003; Roy et al., 2002). Much like high-throughput techniques to capture bound sequence following ChIP, the mRNA can be isolated, converted to cDNA, and either sequenced or hybridized to a microarray. These studies each have their own set of biases. The overlap between two entirely different tissues like muscles and neurons isolated by the same technique can be fairly significant (Fox et al., 2005; Fox et al., 2007). Nevertheless, by comparing datasets from each of the techniques it becomes clear that a statistically significant number of genes are shared between the experiments, though it is by no means a majority of the genes. Depending on the stringency of parameters and calls, a typical tissue has on the order of one to several thousand genes preferentially associated with it at a given time point. This sample of genes represents the last major component of regulatory networks in which we are interested, complementing the cis-regulatory elements, transcription factor binding sites, and transcription factors.

Component interactions

Much of the research performed on transcription factors in multi-cellular organisms gives a very focused understanding of specific aspects of network function. Modern genome-wide techniques can expand the focus to investigate broader aspects of

network function with a single experiment. When reconstructing a network, it is necessary to first identify the components. Described above are some of the techniques that can be used toward this end. Knowing the major components of a transcriptional regulatory network establishes a basic understanding of the network actions. For instance we know that PAL-1 directly activates *hlh-1* and *unc-120* (Lei et al., 2009) to initiate muscle differentiation but is repressed by *mex-3* in the AB lineage (Draper et al., 1996). However, much of this information was gathered under very specific circumstances and may not hold true at other times, in other conditions, or when other genes are involved. For instance, *hnd-1* helps activate muscle differentiation in the C and D lineages (Baugh et al., 2005b), but it has little role in the MS lineage. Instead, *ceh-51* plays that role (Broitman-Maduro et al., 2009). The more limited the data, the more limited the interpretations.

The response of the entire surrounding network in the absence of a transcription factor helps in understanding that gene's role. Rather than focus solely on the necessity of a transcription factor, we can look at what *does* and *does not* depend on the factor. The role of *hlh-1* in the *C. elegans* embryonic body wall muscle differentiation network is useful and approachable for such analysis.

Various independent groups have studied the body wall muscle differentiation network and there exists a decent understanding of the role that *hlh-1* plays within this network (Williams, 1994 #20, Fukushige, 2005 #19, Fukushige, 2006 #18, Chen, 1994 #17, Baugh, 2003 #5, Baugh, 2006 #2). HLH-1 and UNC-120 lie at the foundation of the differentiation network as master-regulators, much like their respective vertebrate orthologs MyoD/myogenin/MRF4/Myf-5 and SRF. Yet the loss of either of these genes

does not halt muscle differentiation (Baugh et al., 2005b; Chen et al., 1992; Fukushige et al., 2006). These transcription factors are believed to directly activate the transcription of numerous genes involved in terminal muscle differentiation – proteins involved in muscle contraction, depolarization, and signalling – such as actins, myosins, calmodulins, calcium channels, receptors, troponins, and tropomyosins. Most of these genes are not exclusively expressed in body wall muscle cells, as nematodes also have pharyngeal and non-striated muscles that possess contractile functions.

So the question arises: why would a dedicated factor such as *hlh-1*, which serves no known purpose other than myogenesis, be conserved across phyla in a network that can still differentiate in its absence? If its function were truly redundant, it should have disappeared over the course of evolution. So what leads to the network seeming to function without the presence of a principal factor?

These transcription factor interactions have been studied in early embryogenesis. When one factor is knocked down, other factors will either rise or fall (Yanai et al., 2008). This reveals factors that activate and repress each other. Early cross-interactions may lead to the network successfully initializing. However, in the absence of *hlh-1*, *unc-120* actually decreases its expression level in early specification (Yanai et al., 2008) and then we find increases its expression level by later differentiation, but only to wild-type levels. Therefore *unc-120* is not compensating for *hlh-1* in terms of expression levels. Nor does this explain the impact of the mutation on terminal network targets. There are several ways that networks activate their differentiation targets. Muscle differentiation in the sea squirt *Ciona* consists of a mix of multiple independent transcription factors binding to independent regulatory elements in some genes and cooperatively binding to a

single element in other genes, with not all factors regulating to all genes (Brown et al., 2007). Nematode muscle has a similar subdivision and varied targeting of genes by network transcription factors. This division is clearly seen between different tissues, as the pharynx and body wall muscles of worms have different myosins, while the non-striated muscles share their myosins with the body wall muscle (MacLeod et al., 1977; Tabara et al., 1996).

Looking later on in embryogenesis, as differentiation commences, reveals a clear story that different targets are activated by multiple factors in various ways. As expected in the *hlh-1* mutant, a large number of genes lose significant levels of expression. However, an even larger number of genes is unaffected by mutation which helps to explain the successful differentiation in the mutant. Several genes appear utterly dependent on *hlh-1*; more genes are only partially dependent, being expressed at lower levels in the mutant; and many genes are completely unaffected by the mutation, clearly being driven if not primarily at least sufficiently by another transcription factor. The troponins illustrate the dynamics well. The troponin *tnc-2* is expressed in non-body wall muscle tissue, thus being entirely independent of *hlh-1*. At the other extreme, *tnt-3* is dependent on HLH-1 binding and loses most expression in the mutant. Two other troponins, *tnt-2* and *tnt-3*, are expressed in both body wall muscle and non-striated muscle, but have no HLH-1 binding and are unaffected by the mutation. Thus there is some division of the targets, with *hlh-1* not targeting some troponins. On top of that, the major troponin target of *hlh-1* is also partially driven by another factor, as it is not silenced in the mutant.

By looking globally, we see the genes *hll-8* and *mls-1* turn on in the mutant animal muscle. Many of their target genes are also turned on exclusively in the mutant muscle. These genes, as part of the non-striated muscle differentiation network, are normally not seen in body wall muscle and represent a compensatory circuit turning on within the mutant. Not quite a shift in fate, their presence reveals more complicated aspects of the network. Rather than adjusting existing components of the body wall muscle system, the network activates new genes from a separate system. We were able to propose the method of network activation by drawing on our data of HLH-1 binding sites from ChIP, novel transcription factor interactions from synthetic phenotypes, and expression data from RNA-seq revealing HLH-1 dependencies. Without this network-wide data we would likely have missed this unexpected reaction to *hll-1* mutation. Whether *hll-8* is the target due to similarities between the muscle systems or binding properties, or similarities between the fate decision pathways is unknown. The result is that the network does not function in and of itself without *hll-1*, but rather activates additional transcription factors not normally functional in muscle to continue with differentiation.

The transcription factor complement of muscle differentiation is very well conserved. However, its method of activating target genes is partially redundant and allows for a considerable degree of flexibility or robustness. At the sub-cellular level, minor fluctuations in expression levels of proteins could potentially have a very dramatic impact on DNA binding levels. A network that can err on the side of extra transcription factor will probably be more fit than a network that cannot. Feed-forward networks, like

the muscle differentiation network, take full advantage of this flexibility. They are examples of brute force, not precision, engineering.

Such a network is not necessarily as efficient as it could be. Rather, a certain degree of overlap works for the system, but the overlap is too complex and intertwined for natural selection to separate the factors. This is a similar phenomenon to the linking of the Hox clusters whose cis-regulatory elements are intercalated. The balance of binding site creation and loss for UNC-120 and HLH-1 is likely in equilibrium, with strong selection for at least one factor to activate a gene, but little consequence as to which factor it is. And because so many factors depend on one or the other, neither factor may be lost.

This intertwining of transcription factors is not always irreversible. Given their similar roles, similar yet distinct binding sites, and shared co-dependencies with *unc-120*, it is possible that *hlh-1* and *hlh-8* may have arisen in the same tissue. In fact, in Cnidarians the orthologs of *hlh-1* and *hlh-8* are both involved in mesoderm and muscle development (Muller et al., 2003; Spring et al., 2000). The balance of these two factors is shifted in different phyla, with *hlh-1*/MRF playing a major role in vertebrates and nematodes and *hlh-8*/*Twist* playing a major role in insects. The crosstalk between the body wall muscle and non-striated muscle networks may be an artifact of a formerly shared regulation.

By using four different varieties of data (cis-regulatory conservation, transcription factor binding, transcription factor catalogue, and target gene catalogue) we have been able to describe gene regulatory network properties in *C. elegans*, primarily in the

myogenic network. The intersect of these four data types has led to the identification of a novel inter-network transcription factor switch and permitted us to propose a mechanism for this interaction. Knowing the pertinent network components makes possible a broad view of network dependence on a single factor, such as *hlf-1*. Only with modern high-throughput technologies can we accurately describe how the network as a whole functions around its transcription factor core. As our knowledge of these networks improves with more advanced techniques and more precise data, even more insight into network function will be tangible.

FIGURES

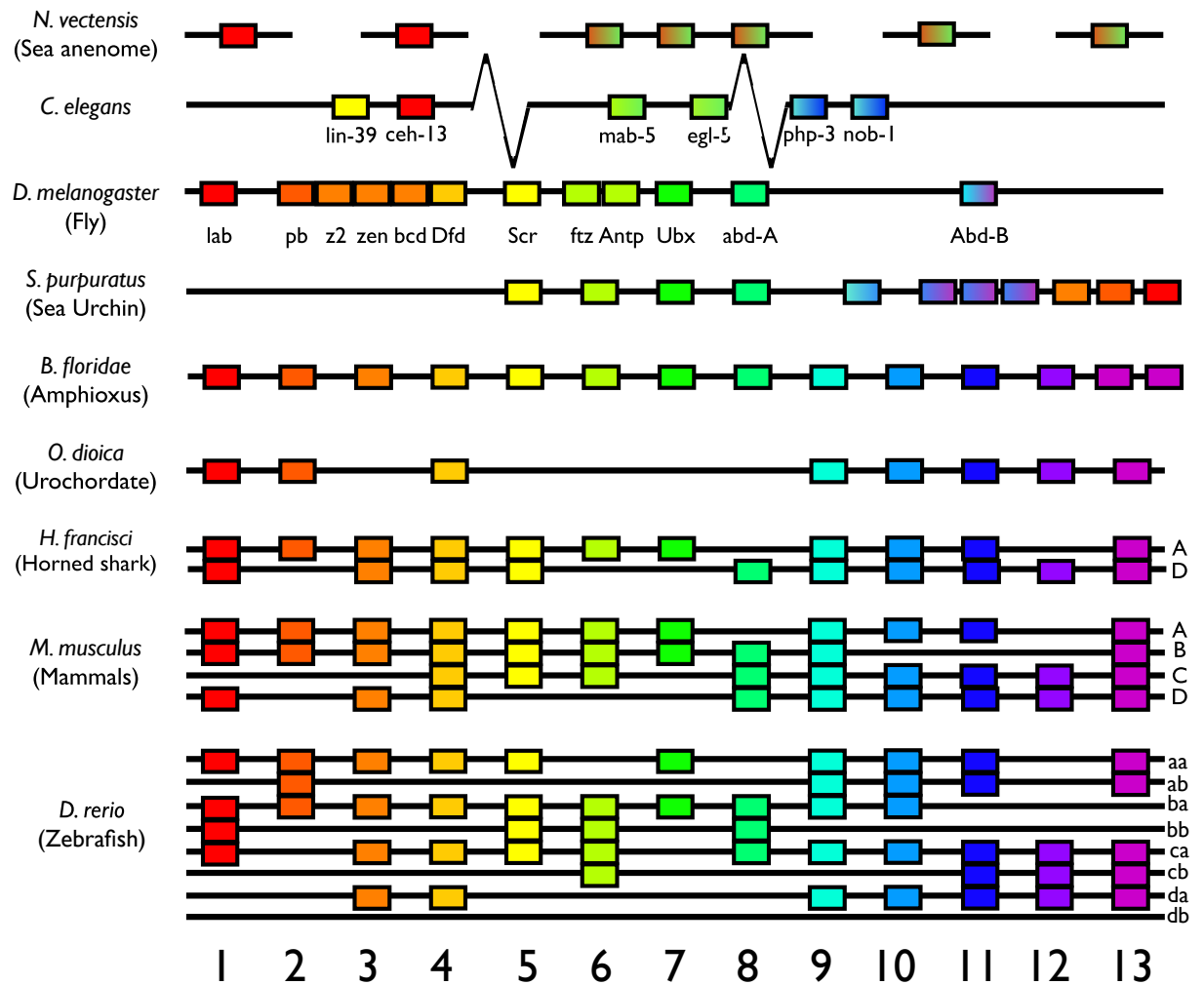


Figure 1: The Hox cluster

All metazoan animals share the Hox cluster, though it is only in the bilateria that it resembles an actual cluster. It has been highly conserved across over 500 million years of evolution and the genes remain joined with one another, possibly due to interlocking cis-regulatory elements. The Hox cluster is interesting due to its central role in development. Determining its regulation is exceedingly difficult. In vertebrates the Hox cluster has undergone duplications, with one duplication in sharks, two duplications in tetrapods, and

three duplications in teleosts. This has made resolving its regulation very difficult. Even in insects, the sheer number of genes that could share regulators is daunting. *C. elegans*, however, only has 6 Hox genes and they are subdivided into three pairs that are separated by megabases of intervening genes and sequence. Therefore it is an ideal organism in which to study the cluster due to its simplicity.

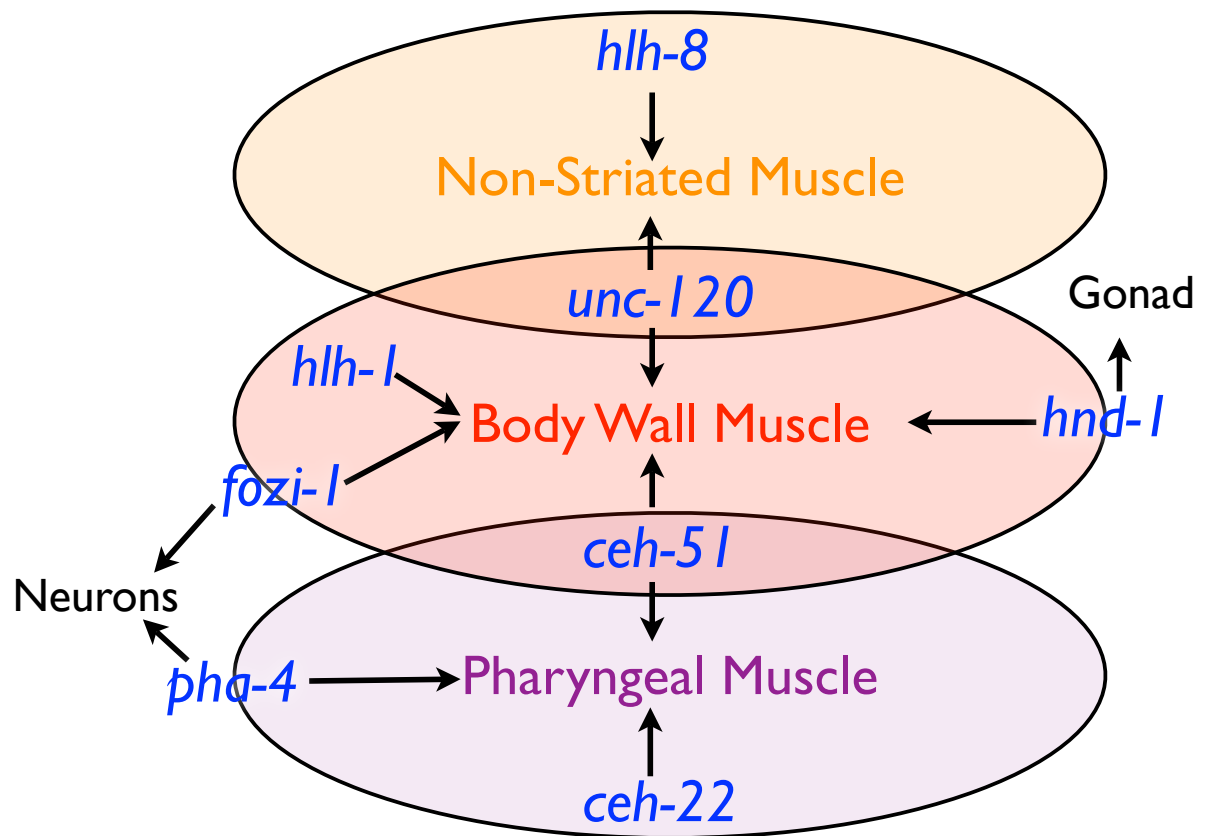


Figure 2: Transcription factor control of myogenesis in the different types of muscle

Different transcription factors control myogenesis in different tissues. The only transcription factors that are completely dedicated appear to be *hlh-1*/MRF, *hlh-8*/Twist, and *ceh-22*/Tinman. *unc-120*/SRF appears to be directly involved in the non-pharyngeal muscles, both the body wall muscle and the non-striated muscles. *ceh-51*/Dlx-1 and *hnd-*

l/Hand1,2 appear to act very early in differentiation but act in very different tissues. *foxi-1* acts during post-embryonic myogenesis as well as in neuron specification.

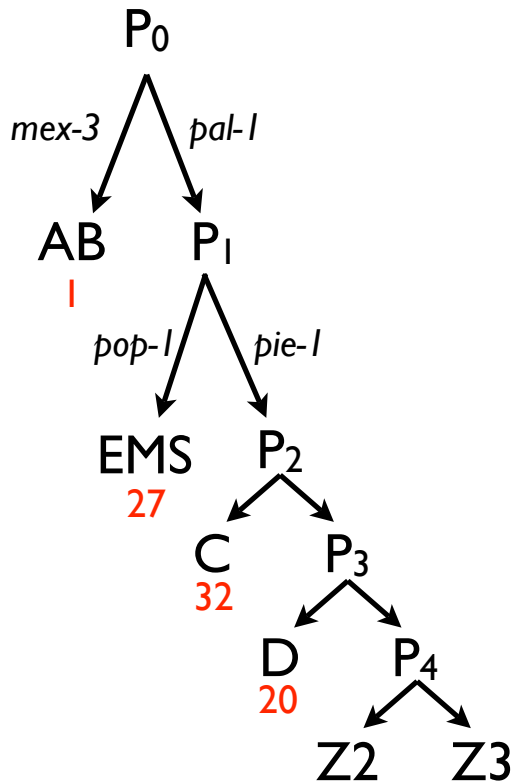


Figure 3: Lineage determination

The embryonic lineages of *C. elegans* depend on certain proteins to direct specification. Knocking out any of these proteins will cause the loss of that particular fate. The absence of *mex-3* prevents the AB lineage from forming; the absence of *pal-1* prevents proper formation of the EMS, C, and D lineages; the absence of *pie-1* prevents the C and D lineages from forming; and the absence of *skn-1* prevents the formation of the EMS lineage. Shown in red are the numbers of body wall muscle cells that normally arise out of each lineage.

REFERENCES

- Aboobaker, A., and Blaxter, M. (2003). Hox gene evolution in nematodes: novelty conserved. *Curr Opin Genet Dev* *13*, 593-598.
- Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L.A., and Rubin, E.M. (2007). Deletion of ultraconserved elements yields viable mice. *PLoS Biol* *5*, e234.
- Amin, N.M., Hu, K., Pruyne, D., Terzic, D., Bretscher, A., and Liu, J. (2007). A Zn-finger/FH2-domain containing protein, FOZI-1, acts redundantly with CeMyoD to specify striated body wall muscle fates in the *Caenorhabditis elegans* postembryonic mesoderm. *Development* *134*, 19-29.
- Ao, W., Gaudet, J., Kent, W.J., Muttumu, S., and Mango, S.E. (2004). Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* *305*, 1743-1746.
- Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* *2*, 28-36.
- Baugh, L.R., Hill, A.A., Claggett, J.M., Hill-Harfe, K., Wen, J.C., Slonim, D.K., Brown, E.L., and Hunter, C.P. (2005a). The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the *C. elegans* embryo. *Development* *132*, 1843-1854.
- Baugh, L.R., Hill, A.A., Slonim, D.K., Brown, E.L., and Hunter, C.P. (2003). Composition and dynamics of the *Caenorhabditis elegans* early embryonic transcriptome. *Development* *130*, 889-900.
- Baugh, L.R., and Hunter, C.P. (2006). MyoD, modularity, and myogenesis: conservation of regulators and redundancy in *C. elegans*. *Genes Dev* *20*, 3342-3346.
- Baugh, L.R., Wen, J.C., Hill, A.A., Slonim, D.K., Brown, E.L., and Hunter, C.P. (2005b). Synthetic lethal analysis of *Caenorhabditis elegans* posterior embryonic patterning genes identifies conserved genetic interactions. *Genome Biol* *6*, R45.
- Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science* *304*, 1321-1325.
- Blackwell, T.K., Bowerman, B., Priess, J.R., and Weintraub, H. (1994). Formation of a monomeric DNA binding domain by Skn-1 bZIP and homeodomain elements. *Science* *266*, 621-628.
- Boffelli, D., Nobrega, M.A., and Rubin, E.M. (2004). Comparative genomics at the vertebrate extremes. *Nat Rev Genet* *5*, 456-465.
- Bowerman, B., Eaton, B.A., and Priess, J.R. (1992). *skn-1*, a maternally expressed gene required to specify the fate of ventral blastomeres in the early *C. elegans* embryo. *Cell* *68*, 1061-1075.
- Braun, T., Bober, E., Rudnicki, M.A., Jaenisch, R., and Arnold, H.H. (1994). MyoD expression marks the onset of skeletal myogenesis in Myf-5 mutant mice. *Development* *120*, 3083-3092.

Broitman-Maduro, G., Owrighi, M., Hung, W.W., Kuntz, S., Sternberg, P.W., and Maduro, M.F. (2009). The NK-2 class homeodomain factor CEH-51 and the T-box factor TBX-35 have overlapping function in *C. elegans* mesoderm development. *Development* *136*, 2735-2746.

Brown, C.D., Johnson, D.S., and Sidow, A. (2007). Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* *317*, 1557-1560.

Brown, C.T., Rust, A.G., Clarke, P.J., Pan, Z., Schilstra, M.J., De Buysscher, T., Griffin, G., Wold, B.J., Cameron, R.A., Davidson, E.H., *et al.* (2002). New computational approaches for analysis of cis-regulatory networks. *Dev Biol* *246*, 86-102.

Brown, C.T., Xie, Y., Davidson, E.H., and Cameron, R.A. (2005). Paircomp, FamilyRelationsII and Cartwheel: tools for interspecific sequence comparison. *BMC Bioinformatics* *6*, 70.

Cameron, R.A., Chow, S.H., Berney, K., Chiu, T.Y., Yuan, Q.A., Kramer, A., Helguero, A., Ransick, A., Yun, M., and Davidson, E.H. (2005). An evolutionary constraint: strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. *Proc Natl Acad Sci U S A* *102*, 11769-11774.

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G.J., Parker, M.H., MacQuarrie, K.L., Davison, J., Morgan, M.T., Ruzzo, W.L., *et al.* (2010). Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Dev Cell* *18*, 662-674.

Chen, L., Krause, M., Draper, B., Weintraub, H., and Fire, A. (1992). Body-wall muscle formation in *Caenorhabditis elegans* embryos that lack the MyoD homolog hlh-1. *Science* *256*, 240-243.

Chen, L., Krause, M., Sepanski, M., and Fire, A. (1994). The *Caenorhabditis elegans* MYOD homologue HLH-1 is essential for proper muscle function and complete morphogenesis. *Development* *120*, 1631-1641.

Clark, S.G., Chisholm, A.D., and Horvitz, H.R. (1993). Control of cell fates in the central body region of *C. elegans* by the homeobox gene *lin-39*. *Cell* *74*, 43-55.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. (2003). Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* *301*, 71-76.

Draper, B.W., Mello, C.C., Bowerman, B., Hardin, J., and Priess, J.R. (1996). MEX-3 is a KH domain protein that regulates blastomere identity in early *C. elegans* embryos. *Cell* *87*, 205-216.

Eddy, S.R. (2005). A model of the statistical power of comparative genome sequence analysis. *PLoS Biol* *3*, e10.

Etchberger, J.F., Lorch, A., Sleumer, M.C., Zapf, R., Jones, S.J., Marra, M.A., Holt, R.A., Moerman, D.G., and Hobert, O. (2007). The molecular signature and cis-regulatory architecture of a *C. elegans* gustatory neuron. *Genes Dev* *21*, 1653-1674.

Fox, J.E., Burow, M.E., McLachlan, J.A., and Miller, C.A., 3rd (2008). Detecting ligands and dissecting nuclear receptor-signaling pathways using recombinant strains of the yeast *Saccharomyces cerevisiae*. *Nat Protoc* *3*, 637-645.

Fox, R.M., Von Stetina, S.E., Barlow, S.J., Shaffer, C., Olszewski, K.L., Moore, J.H., Dupuy, D., Vidal, M., and Miller, D.M., 3rd (2005). A gene expression fingerprint of *C. elegans* embryonic motor neurons. *BMC Genomics* *6*, 42.

Fox, R.M., Watson, J.D., Von Stetina, S.E., McDermott, J., Brodigan, T.M., Fukushige, T., Krause, M., and Miller, D.M., 3rd (2007). The embryonic muscle transcriptome of *Caenorhabditis elegans*. *Genome Biol* 8, R188.

Frasch, M., Chen, X., and Lufkin, T. (1995). Evolutionary-conserved enhancers direct region-specific expression of the murine *Hoxa-1* and *Hoxa-2* loci in both mice and *Drosophila*. *Development* 121, 957-974.

Fukushige, T., Brodigan, T.M., Schriefer, L.A., Waterston, R.H., and Krause, M. (2006). Defining the transcriptional redundancy of early bodywall muscle development in *C. elegans*: evidence for a unified theory of animal muscle development. *Genes Dev* 20, 3395-3406.

Fukushige, T., and Krause, M. (2005). The myogenic potency of HLH-1 reveals wide-spread developmental plasticity in early *C. elegans* embryos. *Development* 132, 1795-1805.

Garcia-Fernandez, J. (2005). The genesis and evolution of homeobox gene clusters. *Nat Rev Genet* 6, 881-892.

Gaudet, J., and Mango, S.E. (2002). Regulation of organogenesis by the *Caenorhabditis elegans* FoxA protein PHA-4. *Science* 295, 821-825.

Gaudet, J., Muttumu, S., Horner, M., and Mango, S.E. (2004). Whole-genome analysis of temporal gene expression during foregut development. *PLoS Biol* 2, e352.

Gonczy, P., Echeverri, C., Oegema, K., Coulson, A., Jones, S.J., Copley, R.R., Duperon, J., Oegema, J., Brehm, M., Cassin, E., *et al.* (2000). Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature* 408, 331-336.

Grove, C.A., De Masi, F., Barrasa, M.I., Newburger, D.E., Alkema, M.J., Bulyk, M.L., and Walhout, A.J. (2009). A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* 138, 314-327.

Guhathakurta, D., Schriefer, L.A., Hresko, M.C., Waterston, R.H., and Stormo, G.D. (2002). Identifying muscle regulatory elements and genes in the nematode *Caenorhabditis elegans*. *Pac Symp Biocomput*, 425-436.

GuhaThakurta, D., Schriefer, L.A., Waterston, R.H., and Stormo, G.D. (2004). Novel transcription regulatory elements in *Caenorhabditis elegans* muscle genes. *Genome Res* 14, 2457-2468.

Haerry, T.E., and Gehring, W.J. (1997). A conserved cluster of homeodomain binding sites in the mouse *Hoxa-4* intron functions in *Drosophila* embryos as an enhancer that is directly regulated by Ultrabithorax. *Dev Biol* 186, 1-15.

Hare, E.E., Peterson, B.K., Iyer, V.N., Meier, R., and Eisen, M.B. (2008). Sepsid even-skipped enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. *PLoS Genet* 4, e1000106.

Harfe, B.D., Branda, C.S., Krause, M., Stern, M.J., and Fire, A. (1998a). MyoD and the specification of muscle and non-muscle fates during postembryonic development of the *C. elegans* mesoderm. *Development* 125, 2479-2488.

Harfe, B.D., Vaz Gomes, A., Kenyon, C., Liu, J., Krause, M., and Fire, A. (1998b). Analysis of a *Caenorhabditis elegans* Twist homolog identifies conserved and divergent aspects of mesodermal patterning. *Genes Dev* 12, 2623-2635.

Hertz, G.Z., and Stormo, G.D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15, 563-577.

Hunt-Newbury, R., Viveiros, R., Johnsen, R., Mah, A., Anastas, D., Fang, L., Halfnight, E., Lee, D., Lin, J., Lorch, A., *et al.* (2007). High-throughput in vivo analysis of gene expression in *Caenorhabditis elegans*. *PLoS Biol* 5, e237.

Hunter, C.P., and Kenyon, C. (1996). Spatial and temporal controls target pal-1 blastomere-specification activity to a single blastomere lineage in *C. elegans* embryos. *Cell* 87, 217-226.

Kassar-Duchossoy, L., Gayraud-Morel, B., Gomes, D., Rocancourt, D., Buckingham, M., Shinin, V., and Tajbakhsh, S. (2004). Mrf4 determines skeletal muscle identity in Myf5:Myod double-mutant mice. *Nature* 431, 466-471.

Kato, M., and Sternberg, P.W. (2009). The *C. elegans* tailless/Tlx homolog nhr-67 regulates a stage-specific program of linker cell migration in male gonadogenesis. *Development* 136, 3907-3915.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241-254.

Kirouac, M., and Sternberg, P.W. (2003). cis-Regulatory control of three cell fate-specific genes in vulval organogenesis of *Caenorhabditis elegans* and *C. briggsae*. *Dev Biol* 257, 85-103.

Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., *et al.* (2005). Combinatorial microRNA target predictions. *Nat Genet* 37, 495-500.

Kuntz, S.G., Schwarz, E.M., DeModena, J.A., De Buyscher, T., Trout, D., Shizuya, H., Sternberg, P.W., and Wold, B.J. (2008). Multigenome DNA sequence conservation identifies Hox cis-regulatory elements. *Genome Res* 18, 1955-1968.

Lei, H., Liu, J., Fukushige, T., Fire, A., and Krause, M. (2009). Caudal-like PAL-1 directly activates the bodywall muscle module regulator hlh-1 in *C. elegans* to initiate the embryonic muscle gene regulatory network. *Development* 136, 1241-1249.

Lemons, D., and McGinnis, W. (2006). Genomic evolution of Hox gene clusters. *Science* 313, 1918-1922.

Li, X.Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C.L., *et al.* (2008). Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol* 6, e27.

Lu, X., and Horvitz, H.R. (1998). lin-35 and lin-53, two genes that antagonize a *C. elegans* Ras pathway, encode proteins similar to Rb and its binding protein RbAp48. *Cell* 95, 981-991.

MacLeod, A.R., Waterston, R.H., Fishpool, R.M., and Brenner, S. (1977). Identification of the structural gene for a myosin heavy-chain in *Caenorhabditis elegans*. *J Mol Biol* 114, 133-140.

Malicki, J., Cianetti, L.C., Peschle, C., and McGinnis, W. (1992). A human HOX4B regulatory element provides head-specific expression in *Drosophila* embryos. *Nature* 358, 345-347.

Maloof, J.N., and Kenyon, C. (1998). The Hox gene lin-39 is required during *C. elegans* vulval induction to select the outcome of Ras signaling. *Development* 125, 181-190.

Mango, S.E., Lambie, E.J., and Kimble, J. (1994). The *pha-4* gene is required to generate the pharyngeal primordium of *Caenorhabditis elegans*. *Development* *120*, 3019-3031.

Mathies, L.D., Henderson, S.T., and Kimble, J. (2003). The *C. elegans* *Hand* gene controls embryogenesis and early gonadogenesis. *Development* *130*, 2881-2892.

McGhee, J.D., Sleumer, M.C., Bilenky, M., Wong, K., McKay, S.J., Goszczynski, B., Tian, H., Krich, N.D., Khattra, J., Holt, R.A., *et al.* (2007). The ELT-2 GATA-factor and the global regulation of transcription in the *C. elegans* intestine. *Dev Biol* *302*, 627-645.

McKay, S.J., Johnsen, R., Khattra, J., Asano, J., Baillie, D.L., Chan, S., Dube, N., Fang, L., Goszczynski, B., Ha, E., *et al.* (2003). Gene expression profiling of cells, tissues, and developmental stages of the nematode *C. elegans*. *Cold Spring Harb Symp Quant Biol* *68*, 159-169.

Michaux, G., Legouis, R., and Labouesse, M. (2001). Epithelial biology: lessons from *Caenorhabditis elegans*. *Gene* *277*, 83-100.

Muller, P., Seipel, K., Yanze, N., Reber-Muller, S., Streitwolf-Engel, R., Stierwald, M., Spring, J., and Schmid, V. (2003). Evolutionary aspects of developmentally regulated helix-loop-helix transcription factors in striated muscle of jellyfish. *Dev Biol* *255*, 216-229.

Oh, S.W., Mukhopadhyay, A., Dixit, B.L., Raha, T., Green, M.R., and Tissenbaum, H.A. (2006). Identification of direct DAF-16 targets controlling longevity, metabolism and diapause by chromatin immunoprecipitation. *Nat Genet* *38*, 251-257.

Okkema, P.G., Harrison, S.W., Plunger, V., Aryana, A., and Fire, A. (1993). Sequence requirements for myosin gene expression and regulation in *Caenorhabditis elegans*. *Genetics* *135*, 385-404.

Olson, E.N., Arnold, H.H., Rigby, P.W., and Wold, B.J. (1996). Know your neighbors: three phenotypes in null mutants of the myogenic bHLH gene MRF4. *Cell* *85*, 1-4.

Ovcharenko, I., Loots, G.G., Nobrega, M.A., Hardison, R.C., Miller, W., and Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. *Genome Res* *15*, 137-145.

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

Pavesi, G., Mauri, G., and Pesole, G. (2004). In silico representation and discovery of transcription factor binding sites. *Brief Bioinform* *5*, 217-236.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., *et al.* (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature* *444*, 499-502.

Peterson, B.K., Hare, E.E., Iyer, V.N., Storage, S., Conner, L., Papaj, D.R., Kurashima, R., Jang, E., and Eisen, M.B. (2009). Big genomes facilitate the comparative identification of regulatory elements. *PLoS ONE* *4*, e4688.

Popperl, H., Bienz, M., Studer, M., Chan, S.K., Aparicio, S., Brenner, S., Mann, R.S., and Krumlauf, R. (1995). Segmental expression of *Hoxb-1* is controlled by a highly conserved autoregulatory loop dependent upon *exd/pbx*. *Cell* *81*, 1031-1042.

Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E.M., Couronne, O., and Pennacchio, L.A. (2006). Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* 16, 855-863.

Rawls, A., Valdez, M.R., Zhang, W., Richardson, J., Klein, W.H., and Olson, E.N. (1998). Overlapping functions of the myogenic bHLH genes MRF4 and MyoD revealed in double mutant mice. *Development* 125, 2349-2358.

Ririe, T.O., Fernandes, J.S., and Sternberg, P.W. (2008). The *Caenorhabditis elegans* vulva: a post-embryonic gene regulatory network controlling organogenesis. *Proc Natl Acad Sci U S A* 105, 20095-20099.

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

Rudnicki, M.A., Braun, T., Hinuma, S., and Jaenisch, R. (1992). Inactivation of MyoD in mice leads to up-regulation of the myogenic HLH gene Myf-5 and results in apparently normal muscle development. *Cell* 71, 383-390.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32, D91-94.

Shim, Y.H., Bonner, J.J., and Blumenthal, T. (1995). Activity of a *C. elegans* GATA transcription factor, ELT-1, expressed in yeast. *J Mol Biol* 253, 665-676.

Sinha, S., Schroeder, M.D., Unnerstall, U., Gaul, U., and Siggia, E.D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in *Drosophila*. *BMC Bioinformatics* 5, 129.

Spith, J., Shim, Y.H., Lea, K., Conrad, R., and Blumenthal, T. (1991). *elt-1*, an embryonically expressed *Caenorhabditis elegans* gene homologous to the GATA transcription factor family. *Mol Cell Biol* 11, 4651-4659.

Spring, J., Yanze, N., Middel, A.M., Stierwald, M., Groger, H., and Schmid, V. (2000). The mesoderm specification factor twist in the life cycle of jellyfish. *Dev Biol* 228, 363-375.

Stone, E.A., Cooper, G.M., and Sidow, A. (2005). Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu Rev Genomics Hum Genet* 6, 143-164.

Stoyanov, C.N., Fleischmann, M., Suzuki, Y., Tapparel, N., Gautron, F., Streit, A., Wood, W.B., and Muller, F. (2003). Expression of the *C. elegans* labial orthologue *ceh-13* during male tail morphogenesis. *Dev Biol* 259, 137-149.

Streit, A., Kohler, R., Marty, T., Belfiore, M., Takacs-Vellai, K., Vigano, M.A., Schnabel, R., Affolter, M., and Muller, F. (2002). Conserved regulation of the *Caenorhabditis elegans* labial/Hox1 gene *ceh-13*. *Dev Biol* 242, 96-108.

Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983). The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev Biol* 100, 64-119.

Tabara, H., Motohashi, T., and Kohara, Y. (1996). A multi-well version of in situ hybridization on whole mount embryos of *Caenorhabditis elegans*. *Nucleic Acids Res* 24, 2119-2124.

Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., and Jones, R.T. (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* 203, 439-455.

Valdez, M.R., Richardson, J.A., Klein, W.H., and Olson, E.N. (2000). Failure of Myf5 to support myogenic differentiation without myogenin, MyoD, and MRF4. *Dev Biol* 219, 287-298.

Wagmaister, J.A., Miley, G.R., Morris, C.A., Gleason, J.E., Miller, L.M., Kornfeld, K., and Eisenmann, D.M. (2006). Identification of cis-regulatory elements from the *C. elegans* Hox gene *lin-39* required for embryonic expression and for regulation by the transcription factors LIN-1, LIN-31 and LIN-39. *Dev Biol* 297, 550-565.

Wang, B.B., Muller-Immergluck, M.M., Austin, J., Robinson, N.T., Chisholm, A., and Kenyon, C. (1993). A homeotic gene cluster patterns the anteroposterior body axis of *C. elegans*. *Cell* 74, 29-42.

Wang, P., Zhao, J., and Corsi, A.K. (2006). Identification of novel target genes of CeTwist and CeE/DA. *Dev Biol* 293, 486-498.

Wang, Y., and Jaenisch, R. (1997). Myogenin can substitute for Myf5 in promoting myogenesis but less efficiently. *Development* 124, 2507-2513.

Waterston, R.H. (1989). The minor myosin heavy chain, *mhcA*, of *Caenorhabditis elegans* is necessary for the initiation of thick filament assembly. *EMBO J* 8, 3429-3436.

Wenick, A.S., and Hobert, O. (2004). Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in *C. elegans*. *Dev Cell* 6, 757-770.

Whittle, C.M., McClinic, K.N., Ercan, S., Zhang, X., Green, R.D., Kelly, W.G., and Lieb, J.D. (2008). The genomic distribution and function of histone variant HTZ-1 during *C. elegans* embryogenesis. *PLoS Genet* 4, e1000187.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434, 338-345.

Yanai, I., Baugh, L.R., Smith, J.J., Roehrig, C., Shen-Orr, S.S., Claggett, J.M., Hill, A.A., Slonim, D.K., and Hunter, C.P. (2008). Pairing of competitive and topologically distinct regulatory modules enhances patterned gene expression. *Mol Syst Biol* 4, 163.

Yuh, C.H., Brown, C.T., Livi, C.B., Rowen, L., Clarke, P.J., and Davidson, E.H. (2002). Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin. *Dev Biol* 246, 148-161.

Zhang, W., Behringer, R.R., and Olson, E.N. (1995). Inactivation of the myogenic bHLH gene MRF4 results in up-regulation of myogenin and rib anomalies. *Genes Dev* 9, 1388-1399.

Zhao, G., Schriefer, L.A., and Stormo, G.D. (2007a). Identification of muscle-specific regulatory modules in *Caenorhabditis elegans*. *Genome Res* 17, 348-357.

Zhao, J., Wang, P., and Corsi, A.K. (2007b). The *C. elegans* Twist target gene, *arg-1*, is regulated by distinct E box promoter elements. *Mech Dev* 124, 377-389.

Zhong, M., Niu, W., Lu, Z.J., Sarov, M., Murray, J.I., Janette, J., Raha, D., Sheaffer, K.L., Lam, H.Y., Preston, E., *et al.* (2010). Genome-wide identification of binding sites defines distinct functions for *Caenorhabditis elegans* PHA-4/FOXA in development and environmental response. *PLoS Genet* 6, e1000848.



Multigenome DNA sequence conservation identifies *Hox cis*-regulatory elements

Steven G. Kuntz, Erich M. Schwarz, John A. DeModena, et al.

Genome Res. 2008 18: 1955-1968 originally published online November 3, 2008

Access the most recent version at doi:[10.1101/gr.085472.108](https://doi.org/10.1101/gr.085472.108)

Supplemental Material

<http://genome.cshlp.org/content/suppl/2008/11/06/gr.085472.108.DC1.html>

References

This article cites 83 articles, 33 of which can be accessed free at:
<http://genome.cshlp.org/content/18/12/1955.full.html#ref-list-1>

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#)

To subscribe to *Genome Research* go to:
<http://genome.cshlp.org/subscriptions>

Methods

Multigenome DNA sequence conservation identifies *Hox cis*-regulatory elements

Steven G. Kuntz,^{1,2} Erich M. Schwarz,¹ John A. DeModena,^{1,2} Tristan De Buysscher,¹ Diane Trout,¹ Hiroaki Shizuya,¹ Paul W. Sternberg,^{1,2,3} and Barbara J. Wold^{1,3}

¹Division of Biology, California Institute of Technology, Pasadena, California 91125, USA; ²Howard Hughes Medical Institute, California Institute of Technology, Pasadena, California 91125, USA

To learn how well ungapped sequence comparisons of multiple species can predict *cis*-regulatory elements in *Caenorhabditis elegans*, we made such predictions across the large, complex *ceh-13/lin-39* locus and tested them transgenically. We also examined how prediction quality varied with different genomes and parameters in our comparisons. Specifically, we sequenced ~0.5% of the *C. brenneri* and *C. sp. 3 PSIOIO* genomes, and compared five *Caenorhabditis* genomes (*C. elegans*, *C. briggsae*, *C. brenneri*, *C. remanei*, and *C. sp. 3 PSIOIO*) to find regulatory elements in 22.8 kb of noncoding sequence from the *ceh-13/lin-39 Hox* subcluster. We developed the MUSSA program to find ungapped DNA sequences with N-way transitive conservation, applied it to the *ceh-13/lin-39* locus, and transgenically assayed 21 regions with both high and low degrees of conservation. This identified 10 functional regulatory elements whose activities matched known *ceh-13/lin-39* expression, with 100% specificity and a 77% recovery rate. One element was so well conserved that a similar mouse *Hox* cluster sequence recapitulated the native nematode expression pattern when tested in worms. Our findings suggest that ungapped sequence comparisons can predict regulatory elements genome-wide.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>) under accession nos. FJ362353–FJ36238.]

Despite knowledge of entire genome sequences, discovering *cis*-regulatory DNA elements remains surprisingly inefficient. In animal genomes, *cis*-regulatory elements are located unpredictably around or within the genes they regulate (Woolfe et al. 2005; Davidson 2006; Pennacchio et al. 2006; Engström et al. 2007). These elements, when dissected further, often prove to be composed of individual transcription factor binding sites that are often very loosely defined (Sandelin et al. 2004). Transgenic analysis *in vivo* is the most definitive way to show that a sequence is regulatory, but it is also the most time consuming and expensive. It is therefore desirable to use other criteria, such as preferential sequence conservation, to identify regions most likely to be functional. To evaluate a strategy for phylogenetic footprinting using four other *Caenorhabditis* species, we dissected the *cis*-regulatory structure of a *Hox* cluster in the nematode *Caenorhabditis elegans* (Fig. 1A).

If two or more species are evolutionarily close enough to show common development and physiology, their genomes are expected to share an underlying gene regulatory network driven by *cis*-regulatory elements with conserved sequences of several hundred base pairs (Tagle et al. 1988; Davidson 2006; Brown et al. 2007; Li et al. 2007). Within a functional *cis*-regulatory element, individual transcription-factor binding sites are generally short (~6–20 bp) with statistical preferences, not strict requirements, for specific bases (Sandelin et al. 2004). Statistical over-representation of such motifs has been useful for identifying transcription-factor binding sites common to coregulated genes

in *C. elegans* (Ao et al. 2004; Gaudet et al. 2004; Wenick and Hobert 2004; Pauli et al. 2006; Etchberger et al. 2007; McGhee et al. 2007; Zhao et al. 2007). However, this approach requires a known set of coregulated genes, a limitation that cross-species genomic comparison methods do not have. The simplest genomic comparison method is all-against-all matching of ungapped sequence windows, which is well suited for finding *cis*-regulatory elements under selective pressure against insertions and deletions (Brown et al. 2002; Cameron et al. 2005). This kind of comparison reveals orientation-independent, one-to-many, and many-to-many relationships, all of which are possible for conserved *cis*-regulatory sequences, yet invisible in standard global alignments. While ungapped comparisons can highlight regulatory regions, they are not expected to resolve individual transcription-factor binding sites within them. However, different prediction biases from sequence conservation versus statistical over-representation can complement one another (Wang and Stormo 2003; Bigelow et al. 2004; Tompa et al. 2005; Chen et al. 2006).

Since purely random pairing of unrelated 100-bp DNA segments typically yields two perfect 6-bp matches (Dickinson 1991), comparing three or more species should identify sequences under selective pressure with greater accuracy than comparing only two (Boffelli et al. 2004; Sinha et al. 2004; Eddy 2005; Stone et al. 2005). This has recently been done for budding yeasts (Cliften et al. 2003; Kellis et al. 2003), *Drosophila* (Stark et al. 2007), and vertebrates (Krek et al. 2005; Xie et al. 2005, 2007; Pennacchio et al. 2006; McGaughey et al. 2008). Vertebrates have many conserved sequences that may be regulatory, but most have unknown functions (Bejerano et al. 2004; Boffelli et al. 2004; Ovcharenko et al. 2005; Ahituv et al. 2007) that are difficult to test in all cell types throughout the life cycle, especially in mammals.

³Corresponding authors.

E-mail woldb@caltech.edu; fax (626) 395-5750.

E-mail pws@caltech.edu; fax (626) 568-8012.

Article published online before print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.085472.108>.

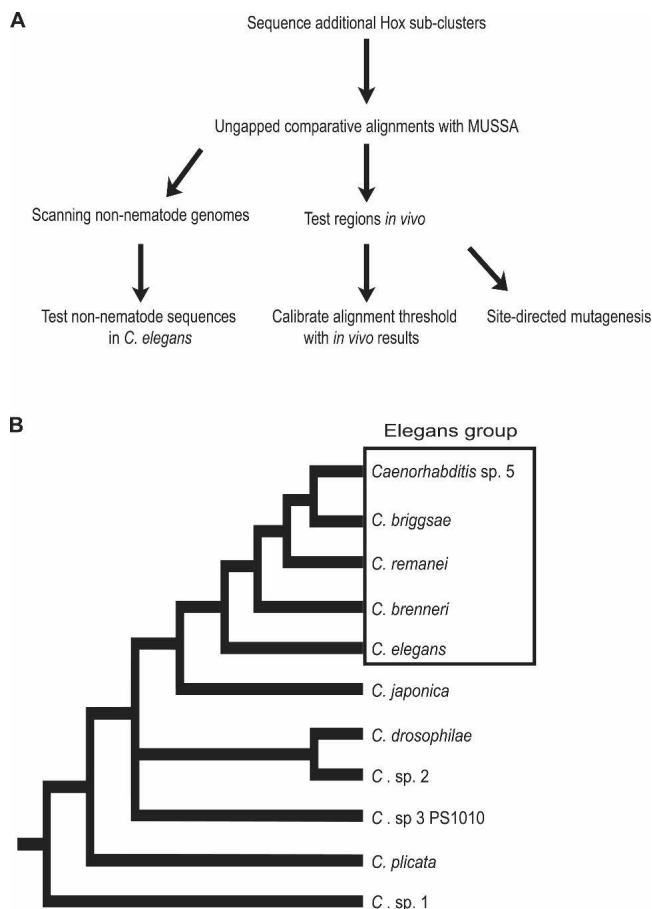


Figure 1. Experimental flow and *Caenorhabditis* phylogeny. (A) The experimental rationale of the project is shown. (B) Phylogeny of nematodes within the *Caenorhabditis* genus from Kiontke et al. (2007). The Elegans group and *C. sp. 3 PS1010* are dealt with in this study.

The nematode *Caenorhabditis elegans* has a compact genome (100 Mb, ~27,000 genes) and body (~1000 somatic cells in adults), which should allow candidate regulatory elements to be tested for function throughout development and across all cell types (Sulston and Horvitz 1977; Kimble and Hirsh 1979; Hillier et al. 2005). Although *C. elegans* is the most familiar *Caenorhabditis* species, others are available for multispecies genomic comparisons (Fig. 1B) (Sudhaus and Kiontke 1996, 2007; Baldwin et al. 1997; Stothard and Pilgrim 2006). Sibling species (the Elegans group, including *C. brenneri*) are difficult to distinguish from *C. elegans* morphologically, save for sex differences (Sudhaus and Kiontke 1996; Kiontke et al. 2004). *C. japonica*, the closest outgroup, shows some morphological differences, but they are relatively minor (Kiontke et al. 2002), while the more distant *C. sp. 3 PS1010* has distinct morphology and behavior (Sudhaus and Kiontke 1996; Cho et al. 2004; Kiontke et al. 2004). Since *C. brenneri* subdivides an evolutionary branch between *C. elegans* and the siblings *C. briggsae* and *C. remanei*, comparisons of its genome with the others might help weed out nonfunctional DNA sequences that had failed to diverge in the sibling species. Comparisons with the more remote *C. sp. 3 PS1010* might define more highly conserved sequences invariant within the *Caenorhabditis*

genus and not simply within the Elegans group. We therefore undertook a pilot project to sequence and analyze ~0.5% of the genomes of *C. brenneri* and *C. sp. 3 PS1010*, including the *Hox* subcluster *ceh-13/lin-39* (Streit et al. 2002; Stoyanov et al. 2003; Sternberg 2005; Wagmaister et al. 2006).

ceh-13 and *lin-39* are a linked pair of *Hox* genes, orthologous to *labial/HOXA1* and *Sex combs reduced/HOXA5*. *Hox* genes, an ancient class of developmental control genes, pose a special challenge to *cis*-regulatory analysis because they are not regulated as isolated loci. Instead, they are found throughout bilateria as conserved multigene clusters encoding paralogous transcription factors that are crucial for development, and that are expressed in complex spatiotemporal patterns requiring intricate transcriptional regulation (Garcia-Fernandez 2005; Lemons and McGinnis 2006). *Hox* genes not only function similarly in disparate animal phyla, but may also be regulated similarly (Malicki et al. 1992; Frasch et al. 1995; Popperl et al. 1995; Haerry and Gehring 1997; Streit et al. 2002; Garcia-Fernandez 2005), although few *cis*-regulatory elements shared by *Hox* clusters of different phyla have actually been found (Haerry and Gehring 1997; Streit et al. 2002).

Nematodes have only a single set of *Hox* genes. Several megabases of DNA and numerous non-*Hox* genes separate the *C. elegans* *Hox* cluster into three subclusters of two genes each: *ceh-13/lin-39*, *mab-5/egl-5*, and *nob-1/php-3* (Supplemental Fig. S1) (Aboobaker and Blaxter 2003). This differs from most vertebrate genomes, which have four or five versions of a single large, unfragmented *Hox* gene cluster (Lemons and McGinnis 2006). Some *Hox* genes have been lost in the *C. elegans* lineage, but all those present have vertebrate and arthropod orthologs (Clark et al. 1993; Maloof and Kenyon 1998; Aboobaker and Blaxter 2003; Stoyanov et al. 2003; Wagmaister et al. 2006). *Cis*-regulation is almost certainly confined within each *C. elegans* subcluster: The *ceh-13/lin-39* subcluster is thus a natural experiment, in which two genes represent a cluster of vertebrate orthologs (Lemons and McGinnis 2006).

The *ceh-13/lin-39* subcluster is vital for much anterior and mid-body development in *C. elegans*, but deciphering its *cis*-regulation has been difficult and remains incomplete. It is large by *C. elegans* standards, with almost 20 kb of intergenic DNA encoding only a single microRNA gene. *ceh-13* is required for both embryonic and postembryonic development; null *ceh-13* mutations are lethal (Brunschwig et al. 1999). In the embryo, *ceh-13* is expressed in the A, D, E, and MS lineages and is required for normal gastrulation (Wittmann et al. 1997). Two upstream regulatory sites have been reported to drive expression in the embryo, one of which also acts in the male tail (Streit et al. 2002; Stoyanov et al. 2003). *Cis*-regulation of post-embryonic *ceh-13* expression, which includes the anterior dorsal hypodermis, anterior bodywall muscle, and ventral nerve cord (Brunschwig et al. 1999), is not yet well understood, especially in tissues where it is coexpressed with *lin-39*. While *lin-39* is dispensable for viability, it is required for normal vulval development, migration of the QR and QL neuroblasts, muscle formation, and specification of VC neurons (Burglin and Ruvkun 1993; Clark et al. 1993; Wang et al. 1993; Clandinin et al. 1997; Grant et al. 2000; McKay et al. 2003). A recent study of the *lin-39* promoter delimited several elements to ~300 bp by generating many transgenic reporter strains without using comparative genomics information; one of these elements was critical for vulval expression (Wagmaister et al. 2006). Our working hypothesis is that the complex expression of the *ceh-13/lin-39* locus arises from the summed actions of independent conserved *cis*-regulatory elements.

We have dissected *ceh-13/lin-39* cis-regulation through comparative genomics, and thus defined parameters likely to be useful for genome-wide analyses. This revealed several known and new regulatory elements, including one with functional similarity in mammalian *Hox* clusters.

Results

DNA sequencing

To enable comparisons to *C. elegans*, 1.1 Mb of genomic sequences from *C. brenneri* and *C. sp. 3* PS1010 were sequenced and assembled (Supplemental Tables S1, S2). This comprised ~0.5% of each genome, assuming genome sizes roughly equal to *C. elegans*. The primary DNA sequence data were generally well assembled; the exception was a set of *C. brenneri* clones covering the *mab-5/egl-5* intergenic region, which may have suffered from high polymorphism found in gonochoristic *Caenorhabditis* species (Graustein et al. 2002).

Sequence comparison

We used MUSSA (multi-species sequences analysis; <http://mussa.caltech.edu>) to find preferentially conserved sequences. MUSSA is a N-way sequence comparison algorithm, generalized from Family Relations (Brown et al. 2002), which integrates similarities among three or more genomes (see Methods). It compares, via sliding window, every frame in each participating sequence with every frame in all other sequences, allowing users to choose a window size and threshold of conservation for ungapped sequence matches (here called "MUSSA matches"). MUSSA produces an orientation-independent map of all one-to-one, one-to-many, and many-to-many transitive matches (Fig. 2). MUSSA matches highlight regions intolerant of insertions and deletions that may contain regulatory elements when found outside coding sequences (Cameron et al. 2005).

A number of parallel lines from visualizing MUSSA matches (at a given threshold of conservation) identified domains of similarity between the sequences, indicating the uniqueness and colinearity of potential regulatory elements (Fig. 2). Noise from repeats and low-complexity DNA sequence tended to create a cross-hatched pattern, reflecting many-to-many alignments that could be eliminated by raising similarity thresholds (Fig. 2A).

We initially performed two-way comparisons using a 30-bp window size, which minimized cross-hatched noise and had been useful in comparing mammalian genomes (T. De Buysscher, unpubl.). In principle, the threshold which gives $P \leq 0.05$ for spurious matches in a 30-bp window should be 19/30 identities in 1 kb of completely random sequence (Brown 2006). Since nonconserved sequence is not actually random, the real P -value must be larger. For thresholds of $\leq 21/30$, we found that cross-hatched connections marred the readout (Fig. 2B), while higher thresholds of $\geq 24/30$ revealed a much sparser set of nearly parallel connections (Supplemental Fig. S2A). As expected, comparisons of three or more genomic sequences allowed clean results at lower thresholds than pairwise comparisons, improving the signal-to-noise ratio (Fig. 2A,C; Supplemental Fig. S2A,B).

Three-way comparison of *ceh-13/lin-39* sequences from *C. elegans*, *C. briggsae*, and *C. brenneri* with 30-bp windows identified several conserved regions (Fig. 2A). In *C. elegans*, the *ceh-13/lin-39* locus includes 19 kb of intergenic sequence and 8 kb of intronic

sequence, of which only ~2% was highlighted in MUSSA matches at a threshold of 24/30 (80%). This 50-fold enrichment was the basis for experimental dissection of the locus. In contrast, comparison of *C. elegans*, *C. briggsae*, and *C. sp. 3* PS1010 revealed substantially fewer MUSSA matches and gained no new alignments across the range of parameters (Fig. 2C; Supplemental Fig. S2C–F). After experimentally testing predicted elements, as reported below, we could re-evaluate the effects of window size and genome numbers, as well as determine the effects of using the *C. remanei ceh-13/lin-39* locus (which was unavailable during the earlier part of our work).

Cis-regulatory elements operating during development are typically composed of multiple binding sites arrayed over several hundred base pairs (Davidson 2006; Li et al. 2007). We expected that not all of these binding sites would be preserved as ungapped sequence blocks. To ensure that our comparison parameters did not omit functional sequences from transgenic assays, we buffered each MUSSA match with 200 bp of flanking DNA on each side. Aligned features located close to each other were grouped into single regions for testing. In this manner, 11 different regions (N1–N11) were predicted to be functional (Fig. 3A). The intervening noncoding regions selected for study (I0–I9), being less conserved, were deemed less likely to be functional (Fig. 3A; Supplemental Table S3) but were also tested transgenically.

Four of the 11 conserved regions corresponded to sequences previously shown to have some function. Region N8 corresponds precisely to the microRNA *mir-231* and its upstream promoter. *mir-231* is expressed from embryonic through adult stages, but its biological role is unknown (Lim et al. 2003). Region N3 drives larval ventral nerve cord expression (pJW8) (Wagmaister et al. 2006); region N9 drives embryonic expression (enh450) (Streit et al. 2002); and a region including element N10 drives larval and male tail expression (271-bp enhancer) (Stoyanov et al. 2003). Because our comparison rediscovered elements of the *ceh-13/lin-39* subcluster previously shown to be important, it seemed likely that the newly defined blocks of similarity would also have biological activities.

Expression in *C. elegans*

We tested nine of the 11 strongly conserved regions, and all 10 intervening weakly conserved regions, for their ability to positively regulate expression; their repressor activity (if any) was not assayed. We did not retest the previously characterized N8 and N10, but did retest N3 and N9 to show that our assays reproduced published expression patterns in our reporter system (a $\Delta pes-10$ basal promoter driving nuclear-localized GFP with an *unc-54* 3' untranslated region [UTR]). Background expression from the reporter is described in the Supplemental material, as are experiments showing that different basal promoters gave identical expression patterns in elements that were retested.

Most conserved regions drove expression in specific cell types (Table 1). In all cases, the described expression pattern was reproducible in multiple independent lines. Despite some spatial and temporal overlap, the expression patterns for each region were unique.

The intronic element N1 drove expression in vulval muscle, starting during the L4 larval stage and continuing through the adult (Fig. 4A). This element was well conserved with two MUSSA matches. Region N2 was expressed in the ventral nerve cord during the L1 larval stage (Fig. 4B). Expression of region N2 was also seen in some P cells and in the neural precursor Q cells, which are

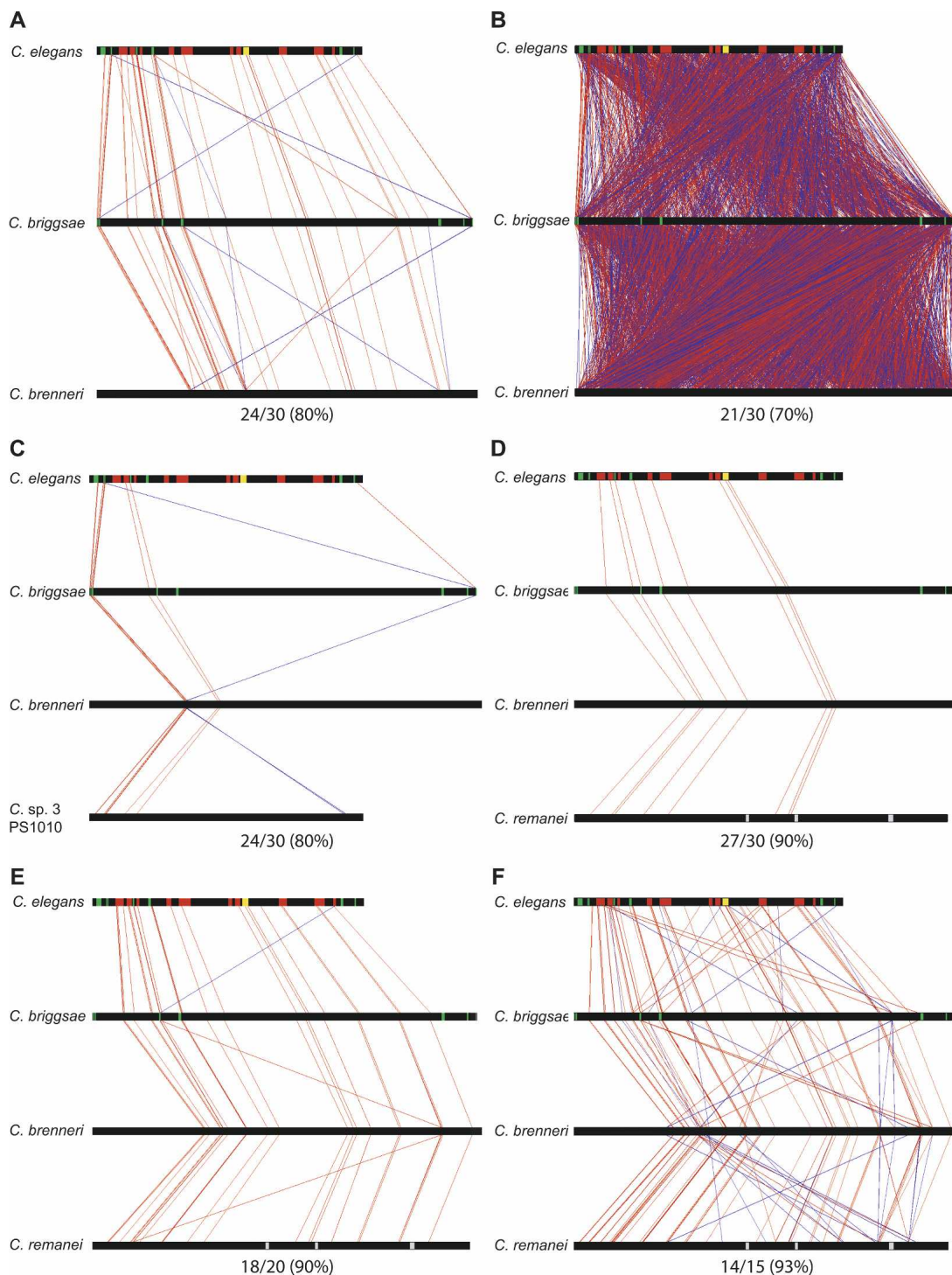


Figure 2. MUSSA comparisons highlighted ungapped sequence matches. Horizontal black bars represent the nematode sequences. The top sequence, *C. elegans*, has green sections for exons (with *lin-39* on the left and *ceh-13* on the right), red sections for each of the N regions, and a yellow section for region N8, which encompasses *mir-231* and its promoter. The vertical lines highlight ungapped sequence MUSSA matches, with red lines for matches facing the same direction and blue lines for reverse-complement matches. The MUSSA matches represent transitive alignments, meaning they match across all sequences compared. (A) At high thresholds the vertical red lines are largely parallel, reflecting predominant colinearity of conserved sequence identified with 80% (24/30) sequence identity for a 30-bp window. As the threshold (identity/window length) decreases, more matches are identified by MUSSA but the noise also increases. (B) At a lower threshold, 70% (21/30), the graph is packed with many lines that cross each other, producing a cluttered, cross-hatched pattern. The number of species being compared may also be varied, giving a range of matches. Comparisons, using a 30-bp window, are shown between *C. elegans*, *C. briggsae*, and *C. brenneri* at 80% (24/30) (A) and *C. elegans*, *C. briggsae*, *C. brenneri*, and *C. sp. 3 PS1010* at 80% (24/30) (C). The window size can also be varied at a constant threshold, as between 27/30 (90%) (D), 18/20 (90%) (E), and 14/15 (93%) (F).

Hox regulatory elements found by genomic conservation

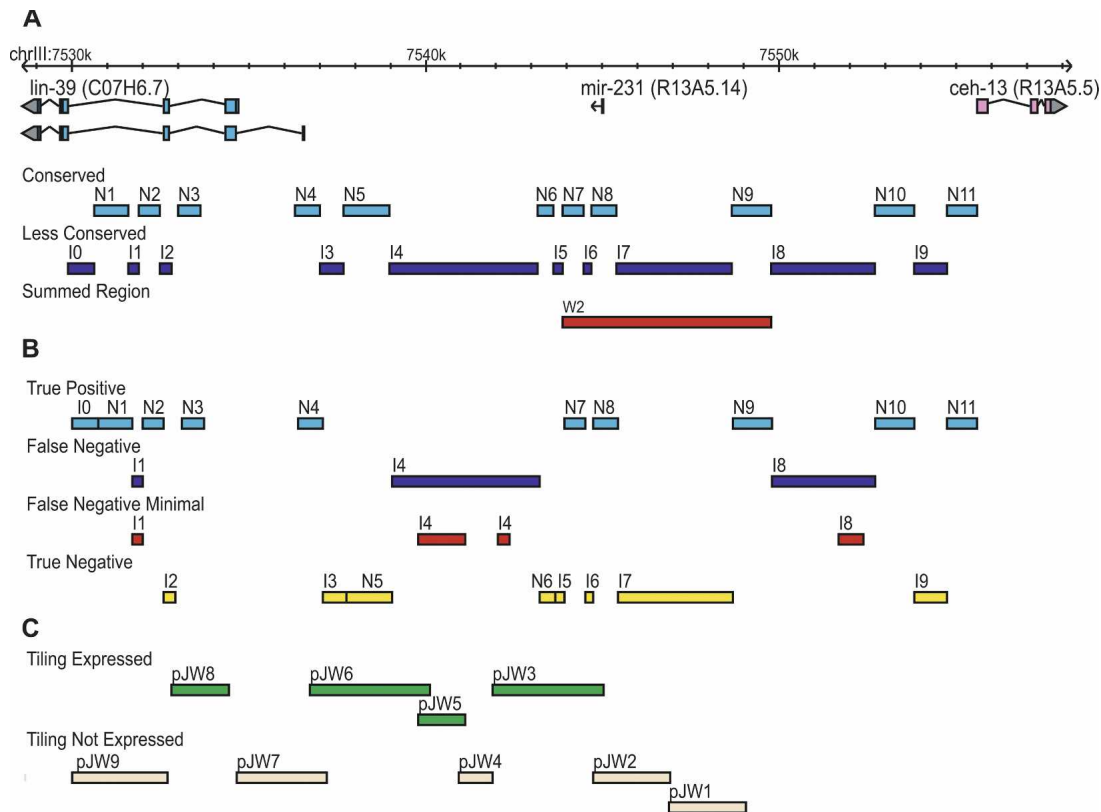


Figure 3. *ceh-13/lin-39* Hox subcluster dissection based on sequence conservation. The *ceh-13/lin-39* Hox locus was dissected into 21 sections for in vivo expression analysis based on the presence of MUSSA matches in a three-way alignment between *C. elegans*, *C. briggsae*, and *C. brenneri*. (A) MUSSA matches were used to identify similar, presumably conserved regions (N regions), which include the sequence match windows, 200 bp of 5' and 3' flanking sequence, and additional sequence for primer selection. The intervening, less-similar regions (I regions) located between the N regions were also tested. A "summed" region (W2) encompassing several component regions is shown as well. (B) With revised parameters of 100% match of 15-bp windows, the regions were repartitioned and true positives, true negatives, and false negatives were identified. The minimal region to recover the observed expression in the false negatives is identified (Streit et al. 2002; Wagmaister et al. 2006). (C) The regions assayed in the tiling analysis from Wagmaister et al. (2006) are shown for comparison, noting which drove expression (green) and which did not (beige).

known to require *lin-39* to regulate proper migration. N2 was also highly conserved: It consisted of two intronic MUSSA matches next to one another in all species except for *C. sp. 3 PS1010*, in which one match was inverted and moved 5' with respect to *lin-39*. N2 occupies the same intron as N1, but is sufficiently separated (by 500 bp in *C. elegans*) to designate N1 and N2 as separate elements. Region N3, identified by one very well-conserved MUSSA match in the first intron of *lin-39*, was expressed in the hypodermal hyp7 cells in the late embryo and early L1 larvae (Fig. 4C) as well as in the V cells, P cells, and ventral nerve cord of the early L1 through L3 larvae. This expression pattern matched and expanded on that previously observed for this region (Wagmaister et al. 2006). Region N4 is in the proximal promoter region of *lin-39*; it drove expression in the ventral mid-body of the early embryo shortly after gastrulation (Fig. 4D). During early larval development N4 also drove expression in V6. Region N7 drove expression in the posterior bodywall muscle cells (Fig. 4E), starting in the late embryo and continuing through adulthood, and in the diagonal and longitudinal muscles of the male tail. Region N9 drove previously reported embryonic expression, along with previously unreported anterior bodywall muscle expression in L4 larvae and adults (Fig. 4F) (Streit et al. 2002). Region N11 was in the proximal promoter region of *ceh-13* and drove expression in the anterior hypodermis

of late embryos (Fig. 4G). Neither N5 nor N6 drove expression; this could be due to the limited conditions (e.g., non-dauer, non-infected, etc.) in which we scored the worms.

Potential regulatory sequences were found for both *ceh-13* and *lin-39*. For conserved regions closer to *ceh-13* (N9 and N11), observed patterns agreed well with expected ones (Wittmann et al. 1997; Brunschwig et al. 1999; Streit et al. 2002). Expression of *lin-39* in the bodywall muscles, intestine, and central body region have all been described and were reproduced, for the most part, by conserved regions closer to *lin-39*: N1–N4 and N7 (Clark et al. 1993; Wang et al. 1993; Maloof and Kenyon 1998; McKay et al. 2003). Furthermore, expression in the anterior midbody is predicted for both transcription factors, meaning that regions N2–N4 could be acting on both genes. Published patterns for both *ceh-13* and *lin-39* may be incomplete, which would account for observed activities beyond those expected.

Each region drove a different expression pattern. The fusion of a large region (W2) that included both N7 and N9 drove expression in both anterior and posterior bodywall muscle, a simple summation of N7 (strictly posterior) and N9 (strictly anterior) expression patterns (Figs. 3A, 4H). It is unknown whether these regions regulate *ceh-13*, *lin-39*, *mir-231*, or all three genes.

We then asked what regulatory activities, if any, resided in the less-conserved regions between our conserved elements.

Table 1. Expression patterns of transgenic worms

Region	Length	Stages	Expression pattern
N1	964	L4-adult	Vulval muscle
N2	605	L1-adult	Ventral nerve cord, Q cell daughters
		L1	P cells, Q cells
N3	630	Embryo-L1	Hyp7
		L1-L3	V cells, P cells, ventral nerve cord
N4	697	Embryo	Ventral midbody
		L1	V6
N5	1297	Embryo-adult	Background (see below)
N6	434	Embryo-adult	Background (see below)
N7	591	Embryo-adult	Posterior bodywall muscle, nerve ring neurons, HSN
N9	1120	L4-adult	Anterior bodywall muscle
N11	819	Embryo	Anterior hypodermis
I0	749	L2-adult	Coelomocytes, anterior ventral nerve cord
		Embryo-L1	V cells, P cells
I1	289	Embryo-adult	Seam cells
I2	311	Embryo-adult	Background (see below)
I3	697	Embryo-adult	Background (see below)
I4	4182	L3	Sex myoblasts
I5	280	Embryo-adult	Background (see below)
I6	216	Embryo-adult	Background (see below)
I7	3270	Embryo-adult	Background (see below)
I8	2906	Embryo	Various
I9	957	Embryo-adult	Background (see below)
W2	5892	L4-adult	Bodywall muscle
pPD107.94		L1-adult	Background (anterior-most and posterior intestine, anterior-most bodywall muscle, anal depressor cell, enteric muscle, excretory cell)
pPD95.75		L1-adult	Background (see above)

The different regions of the *Hox* cluster that drove expression are listed with the corresponding temporal and spatial pattern. Regions with only “background” expression did not drive any unique detectable expression in our assays. Region N10 was previously described and not injected.

Four of the 10 less-conserved regions (I0, I1, I4, and I8) yielded expression apart from the expected background. Region I0 drove expression in the ventral posterior coelomocytes (Fig. 4I) and the two anterior inner longitudinal muscles of the male tail. This element had one MUSSA match that was strongly identified only when the window size was reduced to 15 or 20 bp. Region I1 drove expression in seam cells, starting with the embryo and continuing through to young adults (Fig. 4J). This element had no components strongly identified by MUSSA, with alignments appearing only at relatively low and noisy thresholds. Region I4 drove expression in the sex myoblasts through two cell divisions (Fig. 4K), as previously described by Wagmaister et al. (2006). Although expression was also reported in the Pn.p cells, we did not observe this, perhaps because I4 was not identical to the pJW5 region assayed by Wagmaister et al. (2006). I4 showed no MUSSA matches until a lower threshold of 22/30 bp or a 20-bp window was used, at which point the regions necessary for sex myoblast and ventral hypodermal Pn.p cell expression described by Wagmaister et al. (2006) were identified. Region I8 drove early embryonic expression, as previously reported (Streit et al. 2002). This region had a number of MUSSA matches that appeared as the threshold or window size was lowered.

Testing for sequence necessity

Our DNA regions from the *ceh-13/lin-39* *Hox* subcluster contained not only blocks of ungapped sequence similarity, but also nonconserved sequences in which they were embedded. While these regions clearly drove expression in transgenic worms, our initial survey did not test whether the small conserved matches within them were crucial for regulatory activity. We therefore assayed *in vivo* constructs derived from some of the most highly conserved regions (N1, N2, N3, and N7; Supplemental Tables S3,

S4), in which we mutated the MUSSA match in *C. elegans*. For N7, mutating the MUSSA match completely eliminated expression in the posterior bodywall muscle, showing the match to be needed for regulation (Fig. 5). In contrast, the remaining mutated regions from N1–N3 had the same expression patterns as their respective wild-type constructs. The conserved matches in N1–N3 were themselves dispensable for regulatory activity, yet were closely associated with active regulatory sequences. Our data paralleled previous negative results of Wagmaister et al. (2006) for a point mutation in the N3 region (HP2), which was a possible *Hox* or *Pbx* binding site.

Ultraconserved elements

Hox clusters are evolutionarily ancient, sharing a common origin for all bilaterians (Garcia-Fernandez 2005; Lemons and McGinnis 2006), meaning that some *cis*-regulatory elements in *C. elegans* *ceh-13/lin-39* might be conserved in other bilaterian phyla (Haerry and Gehring 1997; Streit et al. 2002). The following *Hox*-clusters were searched for any possible MUSSA matches to our conserved elements: the single *Hox* clusters of *Drosophila melanogaster*, *Aedes aegypti* (mosquito), *Anopheles gambiae* (mosquito), *Apis mellifera* (honey bee), *Branchiostoma floridae* (lancelet), *Capitella* sp. I (polychaete worm), *Helobdella robusta* (leech), *Lottia gigantea* (snail), *Schistosoma mansoni* (trematode), *Schmidtea mediterranea* (flatworm), and *Tribolium castaneum* (beetle); the four *Hox* clusters of mouse and human; and the seven *Hox* clusters of zebrafish. In each of these genomes we found several matches of uncertain significance. We therefore searched orthologous *Hox* regions for recurrent patterns of MUSSA matches (Fig. 6A). In newly characterized phyla, for which several related genomes had not yet been sequenced, this approach did not help to evalu-

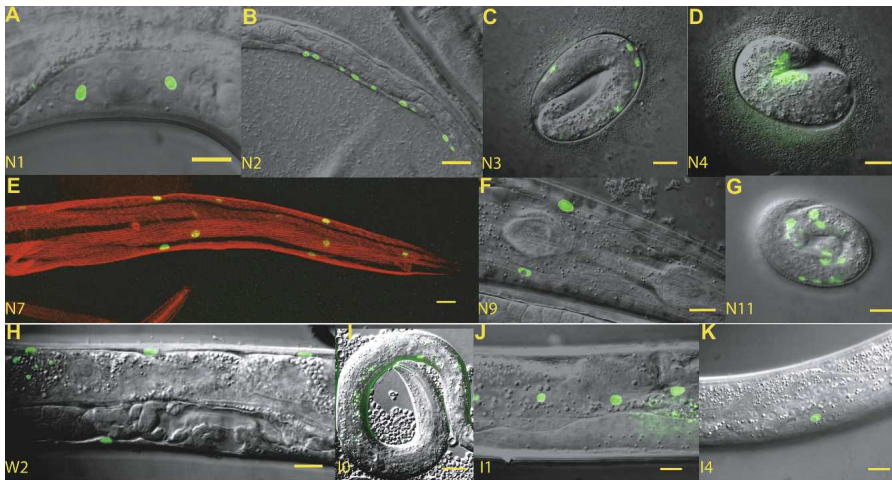


Figure 4. In vivo expression patterns. Many well-conserved and some poorly conserved regions drive independent and reproducible expression. Expression is observed in a variety of tissues that largely agree with published antibody staining for *ceh-13* and *lin-39*. (A) Element N1 directs expression in the L4 to adult vulval muscles. (B) Element N2 directs expression in the late embryo through L2 in the ventral nerve cord and P cells. (C) Element N3 directs expression in late embryonic through L3 hyp7, and in the V cells and P cells soon after hatching. (D) N4 directs expression in cells of the AB lineage in the dorsal mid-body during the comma stage. (E) N7 directs expression in the posterior bodywall muscle in the late embryo through the adult. N8 contains *mir-231* and was not assayed. (F) N9 directs expression in the anterior bodywall muscle in the adult. (G) N11 directs expression in anterior late embryos. (H) W2, a large region spanning N7, N8, and N9, directs expression in both the anterior and posterior bodywall muscles, demonstrating additive coexpression of N7 and N9. (I) I0 directs expression in the posterior ventral coelomocyte. (J) I1 directs expression in the seam cells. (K) I4 directs expression in the SM cells. All scale bars are equal to 10 microns. For background expression from the reporter, see Supplemental material and Supplemental Figure S4.

ate hits; but it was useful in vertebrates and insects, for which many related genomes were available.

In both mouse and human, N3 and N7-like MUSSA matches were paired with each other in the *HOXA* cluster near the *ceh-13* and *lin-39* orthologs, *HOXA1* and *HOXA5*, respectively. Scans of the *HOXA* clusters in dog, opossum, platypus, and frog also revealed this pairing (Fig. 6A). Among the vertebrates alone, sequence conservation was high, indicating that these hits were located in functionally important DNA (Fig. 6B), although these sites had not been previously described. Using a low threshold, the matches showed similarity through nematodes and vertebrates, with the N3-like MUSSA match just 3' of *HOXA1* being more similar (86%) than the N7-like MUSSA match just 5' of *HOXA5* (73%) (Fig. 6C; Supplemental S3A). Similar searches within 11 *Drosophila* species yielded matches highly conserved among insects, but with only low levels of similarity to either nematodes or vertebrates.

To test whether the interphylum similarities revealed functional sequences, we cloned a 700-bp region of mouse *Hox* genomic DNA centered on the mouse N3-like MUSSA match and a 650-bp region centered on the N7-like MUSSA match, each containing local sequence conserved among mammals. We assayed both regions in *C. elegans* transgenes. The mouse N3-like region drove almost the same expression pattern as the *C. elegans* N3 region (Fig. 6D) in hyp7, P cells, V cells, and the ventral nerve cord, with discordant activity in only a few extra anterior hypodermal cells. Whereas *C. elegans* N3 was previously predicted to include a Hox/Pbx autoregulatory site for *lin-39* (Wagmister et al. 2006), the mouse N3-like MUSSA match is found closer to *Hoxa1* (a *ceh-13* ortholog) than to *Hoxa4* (a *lin-39* ortholog). N3 could be a general Hox binding site, or its role may have changed

over time. In contrast, the mouse N7-like region failed to drive the posterior bodywall muscle expression as the *C. elegans* N7 region did, though its background expression level was noticeably increased (Supplemental Fig. S4A).

If N3's similarities between nematodes and vertebrates result from common descent, N3-like matches should exist in other animal phyla. We found co-occurrence of two top-scoring MEME motifs and a MUSSA match in the nematodes, vertebrates, *B. floridae*, *Capitella* sp. I, *H. robusta*, and *S. mansoni* (Supplemental Fig. S3B; Supplemental material). MUSSA comparison of N3-like sequences in nematodes, vertebrates, and *B. floridae* yielded a 70% match, while a comparison of nematodes, vertebrates, *S. mansoni*, and *H. robusta* yielded a 65% match (Supplemental Fig. S3B). These matches encompass deuterostomes, ecdysozoa, and lophotrochozoa—all of the major divisions of bilateria. Thus, we interpret the N3 site to be evolutionarily conserved rather than convergent.

Threshold revision

Having had some success with our initial parameters for ungapped sequence comparison, we then adjusted them empirically and retested them computationally against well-characterized genes in the hope of optimizing our parameters for genome-wide analysis. Initially, nine of the 11 regions (82%) identified by conservation gave expression, while three of the 10 less conserved regions (30%) gave expression; this was promising, but left room for possible improvement. When we tried lower thresholds or smaller windows, MUSSA found matches in some regions that had previously given no hits despite having regulatory activity (and that we had originally classified as false negatives). We therefore optimized the parameter settings and genome combination to achieve the best yield of functional elements while keeping false positives to a minimum (Fig. 2D–F; Supplemental Figs. S2G–L, S5, and S6). A 15-bp window and perfect conservation between *C. elegans*, *C. briggsae*, *C. remanei*, and *C. brenneri* identified MUSSA matches in 77% of all expressing regions with no false positives (Fig. 7A). Using a different window size (14 or 16–30 bp) decreased the resolution and efficiency (see Supplemental material; Supplemental Figs. S5, S6A,B). Including *C. sp. 3* PS1010 sequences adequately selected the top hits, but only at the expense of eliminating many other hits and considerably reducing predictive power (Fig. 7B). Though the four *Elegans* group species together gave the best analysis, inclusion of *C. remanei* masked matches in the I4 region (Supplemental Fig. S2E; see Discussion).

The intervening regions were often much larger than any conserved region. For instance, region I4 was 4.2 kb; however, the subsection of I4 sufficient to drive expression was 1.6 kb (38% of I4) (Wagmister et al. 2006). Likewise, region I8 was 2.9 kb, but expression could be recapitulated with only 0.7 kb within it (24% of I8) (Streit et al. 2002). Thus, the density of regulatory regions

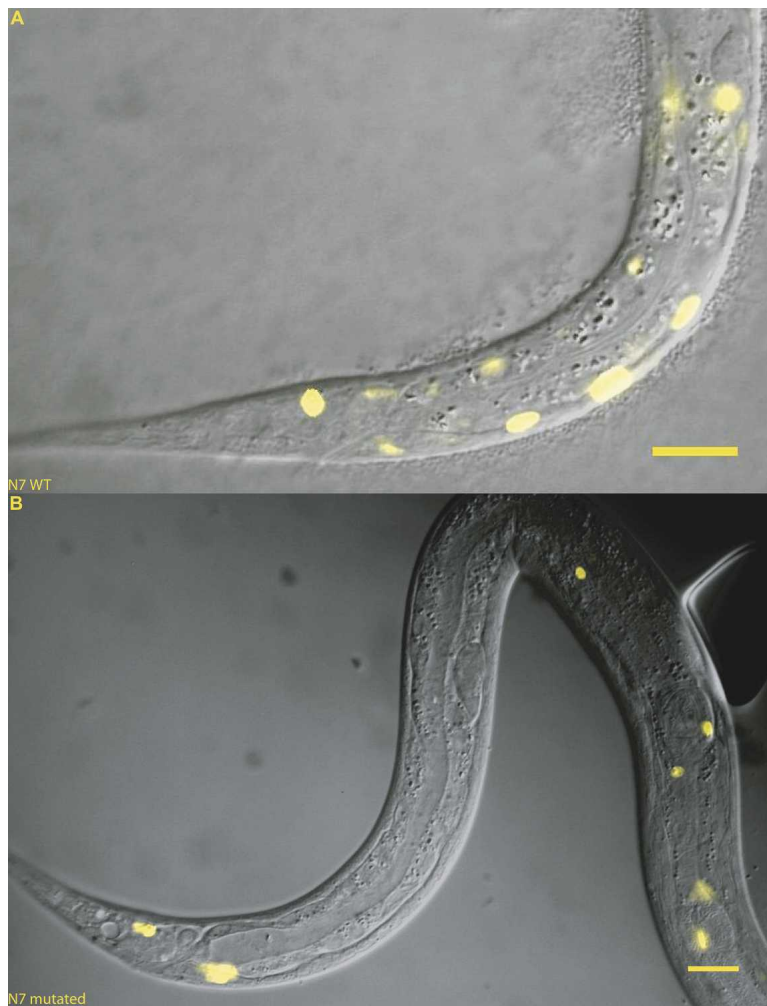


Figure 5. Mutating a conserved window in N7 knocked out expression. Element N7 (592 bp) normally drives expression in the posterior bodywall muscle (A). (B) When the 20-bp MUSSA match was reversed, all expression in the posterior bodywall muscle was abolished. Scale bars, 10 microns.

within nonconserved sequences is probably even lower than our data indicate (Fig. 3B). When compared with tiling, as performed by Wagmaister et al. (2006), conservation-based analysis confers an efficiency advantage, with 100% instead of 40% specificity (Fig. 3C; Wagmaister et al. 2006).

To test whether the revised parameters are useful outside the *Hox* cluster, we analyzed the previously described *C. elegans* genes *hlh-1*, *myo-2*, *myo-3*, and *unc-54* (Okkema et al. 1993; Krause et al. 1994). These were chosen for analysis because their promoter dissections had been screened for expression across all tissues, unlike most studies that identify positive expression in a specific tissue but did not screen for negative activity across other tissues. Using our strict 15-bp threshold and technique of including 200 bp of flanking DNA, all known regulatory elements of the myosin genes *myo-2*, *myo-3*, and *unc-54* (Okkema et al. 1993) were identified with no false positives (Supplemental Fig. S7). For the *hlh-1* locus, two of four regulatory sites (Krause et al. 1994) were recovered at a lower threshold. Therefore, MUSSA predictions were accurate at some non-*Hox* loci, but as in the *Hox* locus itself, some functional elements could not be identified this way.

Discussion

This study found four known and seven new *cis*-regulatory elements in the *ceh-13/lin-39 Hox* subcluster of *C. elegans*, using ungapped sequence conservation across four genomes and verification by transgenic analyses. Remarkably, one conserved element's mouse counterpart recapitulated the native nematode expression pattern. The observed expression patterns generally paralleled those found by prior antibody staining and expression from the parental undissected promoters, suggesting that the union of these *cis*-regulatory elements drives the entire endogenous expression pattern, and that we have identified most *cis*-regulatory regions of *ceh-13/lin-39* (Clark et al. 1993; Wang et al. 1993; Wittmann et al. 1997; Maloof and Kenyon 1998; Brunschwig et al. 1999; Streit et al. 2002; McKay et al. 2003).

For *ceh-13/lin-39*, our first parameters for sequence conservation worked well, even though we later improved them empirically. They identified 11 possible elements, of which nine showed function experimentally, leaving two false positives—a threefold enrichment for functional regulatory elements compared with simple, unselected tiling. With revised parameters, 100% of the computationally identified elements were functional. For these nematode sequences, we found that MUSSA predicted function with highest reliability and resolution when we used windows of 15 bp. Smaller windows gave noisier alignments with poor resolution, while larger windows tended to

miss shorter conserved sequences with regulatory activities. These parameters correctly rediscovered regulatory regions in other well-characterized genes, but made some errors, suggesting additional possible refinements as functional data becomes available at other loci. However, we do not expect that this method, used on its own, will discover all elements. We also expect parameters to change when the set of compared genomes is changed, as we have already found. For instance, the conserved regions for vertebrate *Hox* sequences (e.g., the N3-like mouse region) were much longer than in nematodes, and could be detected at a lower MUSSA threshold with a larger window size. Such differences in sequence conservation might arise from different rates and types of mutations, or from altered selection pressures.

Our aim was to efficiently predict new elements with bona fide biological activity, accepting that this runs the risk of missing some regulatory regions. Nevertheless, correctly identifying even two-thirds of all *C. elegans* regulatory elements with a low false-positive rate, as we did prior to refinement, could significantly advance our knowledge of the worm regulatory genome. Recent uses of sequence constraint in vertebrates have

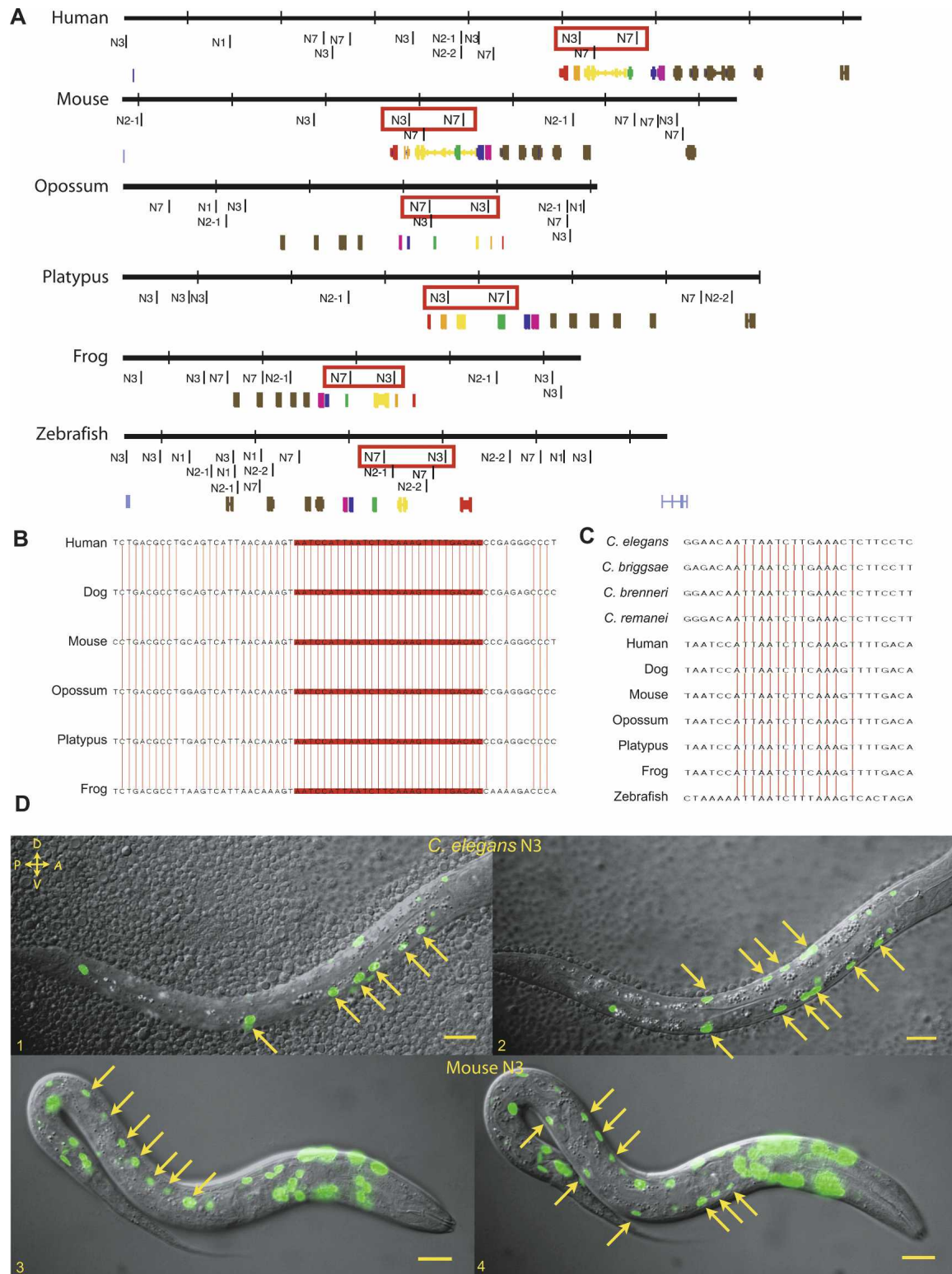


Figure 6. N3 cis-regulatory elements from either nematodes or vertebrates drove expression equivalently. (A) MUSSA analysis was used to identify any ungapped matches between nematodes and various vertebrates. Synteny of two elements, N3 and N7 highlighted by red boxes, suggested the match was not noise. All figures are to the same scale (hash marks represent 50-kb distances), with the regions examined in each case bounded by the next 5' or 3' curated genes on the chromosome. The *Hox* genes are color coded: (red) *HOXA1*, (orange) *HOXA2*, (yellow) *HOXA3*, (green) *HOXA4*, (blue) *HOXA5*, (purple) *HOXA6*. (B) Apparent conservation of N3 among vertebrates was very high, with similarity still at 100% in a 30-bp window. Vertical red lines represent base conservation between all six species. (C) N3 sequences shared 75% identity, using a 20-bp window, across 11 vertebrate and nematode species. (D) A mouse N3-like region drove expression in *C. elegans* that was almost identical to that driven by the *C. elegans* N3 region. Expression is seen in L1 larvae in the V cells on the left (D1, D3), and P cells and hypodermal syncytium on the right (D2, D4). Additional expression is observed in the head with the mouse construct. Scale bars, 10 microns.

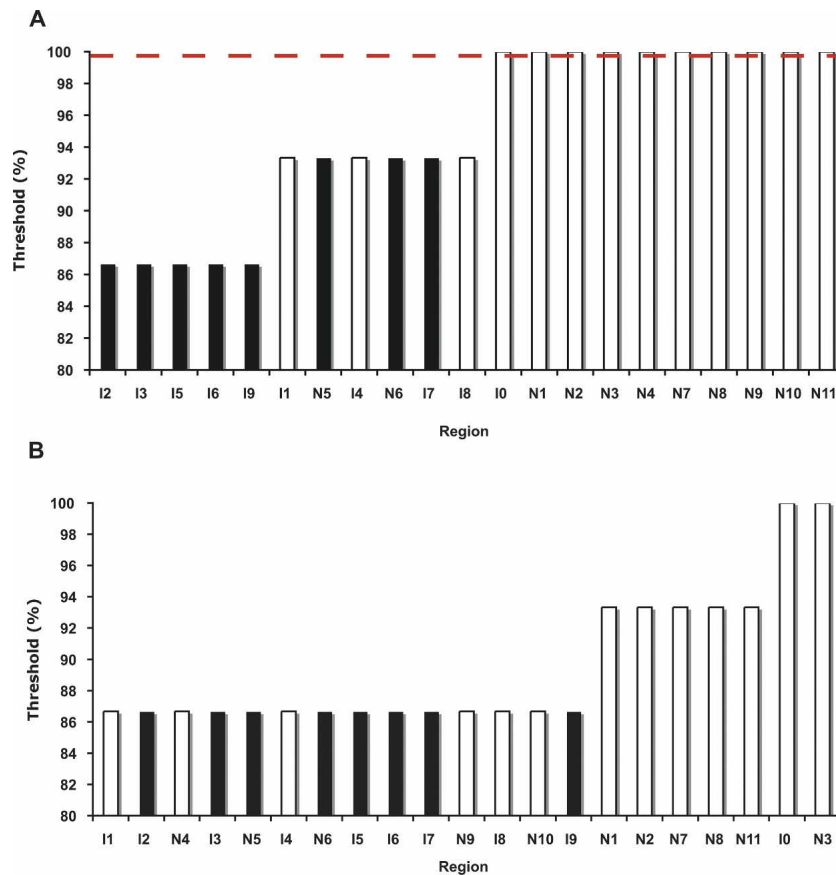


Figure 7. Revising MUSSA parameters for well-conserved regions. (A) A 15-bp window and four-way comparison among *C. elegans*, *C. briggsae*, *C. brenneri*, and *C. remanei* identified the thresholds at which MUSSA matches are observed within a region. Regions capable of driving expression are shown in white and those not capable of driving expression are shown in black. With a threshold of 100%, there is a 77% recovery of expressing regions with perfect specificity. (B) Using five-way comparisons and a 15-bp window among the four above species and *C. sp. 3 PS1010*, the thresholds where conservation was still observed were identified for each element. The predictive power for identifying functional regions is considerably reduced from the four-way comparison.

been less sensitive in finding regulatory elements, perhaps because vertebrates undergo qualitatively different regulation (Pennacchio et al. 2006; McGaughey et al. 2008), although there are many differences, both biological and methodological, between their studies and this one. Only a representative subset of regulatory sites are needed to derive refined, genome-wide motifs in *C. elegans*, as we did with N2-1 (Supplemental material), which can then be statistically correlated with traits of their neighboring genes (Wenick and Hobert 2004; Mortazavi et al. 2006; Etchberger et al. 2007).

If a given regulatory element is mutated or fragmented in some species, comparing it with different sets of related species can still allow detection of that element. Such regulatory mutations are known to be responsible for subtle evolutionary changes in the salt resistance and excretory canal phenotypes of *C. elegans*, which have diverged from the ancestral phenotypes retained in *C. briggsae* and *C. brenneri* (Wang and Chamberlin 2004). The most striking difference in conservation we observed was between *Elegans* group species and the outlying *C. sp. 3 PS1010*. Four-way comparison of *C. elegans*, *C. briggsae*, *C. brenneri*, and *C. remanei* predicted the most regulatory elements, many of which could only be detected in *C. sp. 3 PS1010* with much lower

and noisier thresholds. Although all regions identified with *C. sp. 3 PS1010* drove expression, there was no added benefit from this comparison; rather, it increased the false-negative rate. Similarly, neither *lin-3* nor *lin-11* in *C. sp. 3 PS1010* had the organization or the sequence motifs of the genes in the *Elegans* group species (Supplemental Fig. S8; Supplemental Table S5). Additional *Caenorhabditis* genomic sequences should clarify which parts of the *C. elegans* genome encode species- or group-specific traits.

The regulatory organization of the *ceh-13/lin-39* locus appears to be modular, with each regulatory element functioning independently in transgenes: The expression output of two elements on a single DNA fragment (N7 and N9 on W2) or of four cojoined elements (N1, N2, N3, and N7) matched the sum of their individual activities. Nevertheless, the linear order of conserved elements across the *ceh-13/lin-39* locus has been conserved between the different *Caenorhabditis* species, including the relatively distant *C. sp. 3 PS1010*, suggesting that element order is under selective pressure. Among the elements, there is also potential for some functional redundancy, as has been noted in mammals (e.g., Ahituv et al. 2007). *ceh-13*, for example, is expressed in the larval ventral nerve cord (Brunschwig et al. 1999) and three different elements drive expression there.

Multiple regulatory elements distributed throughout large introns and flanking sequences control many metazoan genes expressed in complex spatiotemporal patterns (Woolfe et al. 2005; Davidson 2006; Pennacchio et al. 2006) and *ceh-13/lin-39* follows this trend. Only two of the nine expressing regions were located within the proximal 2-kb promoter sequences of *ceh-13* or *lin-39*, and four were in *lin-39* introns. We did not assay for the effect that these regions had on *ceh-13*, *lin-39*, or *mir-231* expression. Other examples of distal elements in *C. elegans* include remote regulation of *ceh-10* and *osm-9* (Colbert et al. 1997; Wenick and Hobert 2004).

Conservation analysis helped define elements without inadvertently splitting them, a hazard in blind deletion analysis. Moreover, it may have freed elements from inhibitory sequences, as we found that some large segments were less active when assayed than their subdomains. The entire second intron of *lin-39* yielded no expression in a prior study (Wagmaister et al. 2006), but we identified four different active *cis*-regulatory elements (N1, N2, I0, and I1) by subdividing the region. One possibility is that poorly conserved DNA separating *ceh-13/lin-39* elements harbors hidden regulatory functions that our assay misses, such as repression. The basal promoter construct we used to screen for *in vivo* enhancer activity is not expected to detect isolated transcriptional silencers or insulators. This could explain moderately conserved but inactive regions, as might enhancers

dependent on untested culture conditions or promoter-specific interactions with regulatory elements (Wenick and Hobert 2004; Etchberger et al. 2007).

Although large regions can be split into smaller functional components (such as the W2 region dividing into N7, N8, and N9, and the *lin-39* intron dividing into N1, N2, I0, and I1), further dissection of functional elements might simply disrupt them, yielding weak and variable expression. This has been observed for *ceh-13* male tail expression when multiple sites within N10 were mutated (V. Wegewitz and A. Streit, pers. comm.).

Biologically relevant sequence motifs often appear in or near the best-conserved regions, even if the MUSSA matches themselves are not essential for regulatory activity. For instance, two conserved MUSSA matches <200-bp apart identify the element N9; but a known motif that is not part of either conserved window is located next to them, and is necessary for proper regulatory function (Supplemental Fig. S9A). In four of five mutageneses, changing just one conserved feature had little effect, which is consistent with functional redundancy often seen in multi-site regulatory elements. Our assays used injected transgenes, for which multiple copies generally exist of a cloned reporter (Mello and Fire 1995); this might have provided a relaxed context for gene expression, tolerating the loss of "redundant" sites actually required in vivo. A site that subtly controls the quantity or spatiotemporal pattern of gene activity could easily lack an observable impact on GFP expression. Thus, it is important to test not only conserved sequences for regulatory activity, but the sequences near them.

The apparent conservation of N3 and N7 regions across phyla suggests that they predate the divergence of bilateria. Although mouse N7 was not active in the cross-phylum assay, the mouse N3-like region was strikingly positive and contains a potentially autoregulatory Hox/Pbx binding site. To test regulatory elements for functional conservation between different animal phyla, *Drosophila* enhancers and promoters have been compared with those of *C. elegans* and mammals: This generally involved isolating an enhancer or promoter with a known expression pattern in a donor organism, and testing it transgenically for similar expression in a second, distantly related organism (Malicki et al. 1992; Frasch et al. 1995; Popper et al. 1995; Haery and Gehring 1997; Streit et al. 2002; Ruvinsky and Ruvkun 2003). With nematode and mouse N3 regions, we instead tested the donor enhancer for activity equivalent to that already defined for its ortholog in the recipient species. This provides an alternative for comparisons over very long evolutionary distances, across which anatomical similarities may not be obvious. Moreover, additional MEME motifs, one of which may have been independently identified in mammals (as LM115 and LM171 of Xie et al. [2007]) (Supplemental Results), are shared by the vertebrate and nematode sequences. Based on these in vivo data and computational analyses, we consider N3 a pan-phyletic regulatory sequence. Such sequences may be rare, and only present in the most ancient regulatory loci, such as the ParaHox or NK clusters (Garcia-Fernandez 2005).

Methods

General methods and strains

We obtained *Caenorhabditis elegans*, *C. brenneri* CB5161, and *C. sp. 3 PS1010* from the CGC strain collection and cultured them on OP50 at 20°C, using methods standard for *C. elegans* (Sulston and Hodgkin 1988). *unc-119(ed4)* hermaphrodites were

microinjected with a mixture of 60 ng/μL *unc-119* vector, 12 ng/μL unpurified fusion product, and either 100 ng/μL pBluescript or 100 ng/μL digested genomic DNA to generate transgenic animals (Mello and Fire 1995; Kelly et al. 1997). All noted expression patterns were observed in two or more independent transgenic lines. In nonexpressing lines, at least 16 hermaphrodites from three independent lines (each line driving background GFP to guarantee GFP's functionality) were observed at each stage (early embryos, late embryos, L1–L4 larvae, young adults, and mature adults) with 100× magnification; males and dauers were observed for some, but not all, reporter lines.

DNA preparation

DNA was prepared by standard methods (Sulston and Hodgkin 1988). pEpiFos-5 (Epicentre), based on pBeloBAC11 (Birren et al. 1999), was used as the fosmid library vector. Fosmid sequences were shotgun sequenced and assembled into contigs by the Department of Energy's Joint Genome Institute at Walnut Creek (<http://www.jgi.doe.gov/sequencing/protocols>).

Sequence analysis

Sequence contigs from JGI were initially linked by BLASTN (Korf et al. 2003) and then merged with the *revseq* and *megamerger* functions of EMBOSS (Olson 2002). Our *C. brenneri* data had 22 genomic contigs, totaling 680,633 nucleotides (Supplemental Table S1). Our *C. sp. 3 PS1010* data had seven genomic contigs, totaling 417,129 nucleotides (Supplemental Table S2). Gene predictions were made with Twinscan 3.5 running in single-species mode with *C. elegans* parameters (Wei et al. 2005); predicted protein sequences were extracted with BioPerl (Stajich et al. 2002). *C. brenneri* and *C. sp. 3 PS1010* protein sequences were tested for orthology against one another and against the protein-coding gene sets of *C. elegans*, *C. briggsae*, and *C. remanei* (from the WS170 release of WormBase) with OrthoMCL 1.3 (Li et al. 2003). Inferred ortholog groups were considered specific (i.e., unique) if they contained only one *C. elegans* gene, and only one gene from either *C. briggsae* or *C. remanei*. Our *C. brenneri* contigs encode 141 predicted proteins of ≥100 residues in length, of which 88 have unique *C. elegans* orthologs (Supplemental Table S1). Our *C. sp. 3 PS1010* contigs encode 86 predicted ≥100-residue proteins, 68 with *C. elegans* orthologs (Supplemental Table S2). SVG genomic sequence images were generated by GBrowse for nematodes and vertebrates at the Wormbase (<http://www.wormbase.org>) and UCSC Genome Browser (<http://genome.ucsc.edu>) websites.

MUSSA (multiple species sequences analysis) (<http://mussa.caltech.edu>), a program written in C++ with a Python controlled user interface, was used to identify evolutionarily conserved sequences. MUSSA uses N-way transitivity (all-against-all) so that only windows passing the selected similarity threshold across all species are reported as alignments. No sequences were repeat-masked in the comparisons performed here, though use of MUSSA in other phyla may benefit from masking as a preprocessing step (T. De Buysscher, D. Trout, and B.J. Wold, unpubl.).

For regulatory element dissection in the *ceh-13/lin-39* cluster, published sequences from *C. elegans*, *C. briggsae*, and *C. remanei* (<http://www.wormbase.org>) were used with novel sequences from *C. brenneri* and *C. sp. 3 PS1010*. The *mab-5/egl-5 Hox* cluster comparisons used sequences from *C. elegans*, *C. briggsae*, and *C. remanei*. Additional comparisons with non-nematodes used sequences from all of each organism's available *Hox* clusters (<http://www.ensembl.org>; <http://genome.ucsc.edu>; <http://www.genedb.org/genedb/smanson>; <http://racex00.tamu.edu>; and <http://genome.jgi-psf.org>). Known regulatory regions of

non-*Hox* genes were linked from *C. elegans* to other species using MUSSA.

MEME

Multiple EM for Motif Elicitation (MEME) v3.5.4 was used to identify nonaligned motifs shared by different animal phyla (<http://meme.sdsc.edu/meme>) (Bailey and Elkan 1994). MEME motifs from the N3 element were tested for similarities to previously published genomic motifs by examining two 14-nt human sequences with up to two mismatches against JASPAR CNE (Byrne et al. 2007; Xie et al. 2007).

Transgene design and construction

PCR fusions were generated using standard protocols, essentially as in Hobert (2002). Genomic DNA and the cosmids R13A5 and C07H6 (from A. Fraser and R. Shownkeen at the Sanger Institute) were used as sequence templates. The Fire Lab Vector pPD107.94 was used as the template for the *Δpes-10::4X-NLS::eGFP::lacZ::unc-54* sequence (Mello and Fire 1995). The Fire Lab Vector pPD95.75 was used as the template for the “promoterless” eGFP::*unc-54* sequence (Etchberger and Hobert 2008), used as a control in four constructs to demonstrate identical expression patterns under different basal promoters. Mutation primers were used to mutate target sites in plasmids. The mutated and sequenced enhancers were fused to Fire Lab Vector pPD122.53, where GFP was replaced with YFP, to give a *Δpes-10::4X-NLS::YFP::unc-54*. GFP was replaced with CFP for unmutated controls. We mutated conserved sequences by reversal, not reverse complementation; such reversal maintained the base content, but was expected to destroy any sequence-specific binding of transcription factors. Complete methods are described in the Supplemental material.

Acknowledgments

We dedicate this study to the memory of E.B. Lewis, who pioneered the analysis of *Hox* clusters at Caltech. We thank C.T. Brown for discussions, N. Mullaney for work on an early version of MUSSA, E. Moon for aid in fosmid library construction, and E. Rubin and his colleagues at the DOE JGI for fosmid sequencing and assembly. We thank L.R. Baugh, C.T. Brown, C. Dalal, J. Green, M. Kato, K. Kiontke, A. Mortazavi, A. Seah, and B. Williams for comments on the manuscript. Some nematode strains used in this work were provided by the *Caenorhabditis* Genetics Center, which is funded by the NIH National Center for Research Resources (NCRR). Unpublished metazoan genomic sequences were generously provided by the DOE JGI and GeneDB. This work was supported by grants from DOE to B.J.W. and P.W.S., from NASA to B.J.W., from NIH to B.J.W., and from the HHMI, with which P.W.S. is an Investigator.

References

Aboobaker, A. and Blaxter, M. 2003. Hox gene evolution in nematodes: Novelty conserved. *Curr. Opin. Genet. Dev.* **13**: 593–598.

Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L.A., and Rubin, E.M. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* **5**: e234. doi: 10.1371/journal.pbio.0050234.

Ao, W., Gaudet, J., Kent, W.J., Muttumu, S., and Mango, S.E. 2004. Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* **305**: 1743–1746.

Bailey, T.L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**: 28–36.

Baldwin, J.G., Frisse, L.M., Vida, J.T., Eddleman, C.D., and Thomas, W.K. 1997. An evolutionary framework for the study of developmental

evolution in a set of nematodes related to *Caenorhabditis elegans*. *Mol. Phylogenet. Evol.* **8**: 249–259.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. 2004. Ultraconserved elements in the human genome. *Science* **304**: 1321–1325.

Bigelow, H.R., Wenick, A.S., Wong, A., and Hobert, O. 2004. CisOrtho: A program pipeline for genome-wide identification of transcription factor target genes using phylogenetic footprinting. *BMC Bioinformatics* **5**: 27. doi: 10.1186/1471-2105-5-27.

Birren, B., Mancino, V., and Shizuya, H. 1999. Bacterial artificial chromosomes. In *Genome analysis: A laboratory manual* (eds. B. Birren et al.), pp. 241–295. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Boffelli, D., Nobrega, M.A., and Rubin, E.M. 2004. Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.* **5**: 456–465.

Brown, C.T. 2006. “Tackling the regulatory genome.” Ph.D. thesis, California Institute of Technology, Pasadena.

Brown, C.T., Rust, A.G., Clarke, P.J., Pan, Z., Schilstra, M.J., De Buysscher, T., Griffin, G., Wold, B.J., Cameron, R.A., Davidson, E.H., et al. 2002. New computational approaches for analysis of cis-regulatory networks. *Dev. Biol.* **246**: 86–102.

Brown, C.D., Johnson, D.S., and Sidow, A. 2007. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* **317**: 1557–1560.

Brunschwig, K., Wittmann, C., Schnabel, R., Burglin, T.R., Tobler, H., and Muller, F. 1999. Anterior organization of the *Caenorhabditis elegans* embryo by the labial-like Hox gene *cel-13*. *Development* **126**: 1537–1546.

Byrne, J.C., Valen, E., Tang, M.H., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A. 2007. JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update. *Nucleic Acids Res.* **36**: D102–D106.

Burglin, T.R. and Ruvkun, G. 1993. The *Caenorhabditis elegans* homeobox gene cluster. *Curr. Opin. Genet. Dev.* **3**: 615–620.

Cameron, R.A., Chow, S.H., Berney, K., Chiu, T.Y., Yuan, Q.A., Kramer, A., Helguero, A., Ransick, A., Yun, M., and Davidson, E.H. 2005. An evolutionary constraint: Strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. *Proc. Natl. Acad. Sci.* **102**: 11769–11774.

Chen, N., Mah, A., Blacque, O.E., Chu, J., Phgora, K., Bakhroum, M.W., Newbury, C.R., Khattra, J., Chan, S., Go, A., et al. 2006. Identification of ciliary and ciliopathy genes in *Caenorhabditis elegans* through comparative genomics. *Genome Biol.* **7**: R126. doi: 10.1186/gb-2006-7-12-r126.

Cho, S., Jin, S.W., Cohen, A., and Ellis, R.E. 2004. A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. *Genome Res.* **14**: 1207–1220.

Clandinin, T.R., Katz, W.S., and Sternberg, P.W. 1997. *Caenorhabditis elegans* HOM-C genes regulate the response of vulval precursor cells to inductive signal. *Dev. Biol.* **182**: 150–161.

Clark, S.G., Chisholm, A.D., and Horvitz, H.R. 1993. Control of cell fates in the central body region of *C. elegans* by the homeobox gene *lin-39*. *Cell* **74**: 43–55.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**: 71–76.

Colbert, H.A., Smith, T.L., and Bargmann, C.I. 1997. OSM-9, a novel protein with structural similarity to channels, is required for olfaction, mechanosensation, and olfactory adaptation in *Caenorhabditis elegans*. *J. Neurosci.* **17**: 8259–8269.

Davidson, E.H. 2006. *The regulatory genome: Gene regulatory networks in development and evolution*. Academic Press, San Diego, CA.

Dickinson, W. 1991. The evolution of regulatory genes and patterns in *Drosophila*. *Evol. Biol.* **25**: 127–173.

Eddy, S.R. 2005. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3**: e10. doi: 10.1371/journal.pbio.0030010.

Engström, P.G., Ho Sui, S.J., Drivenes, O., Becker, T.S., and Lenhard, B. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.* **17**: 1898–1908.

Etchberger, J.F. and Hobert, O. 2008. Vector-free DNA constructs improve transgene expression in *C. elegans*. *Nat. Methods* **5**: 3. doi: 10.1038/nmeth0108-3.

Etchberger, J.F., Lorch, A., Sleumer, M.C., Zapf, R., Jones, S.J., Marra, M.A., Holt, R.A., Moerman, D.G., and Hobert, O. 2007. The molecular signature and cis-regulatory architecture of a *C. elegans* gustatory neuron. *Genes & Dev.* **21**: 1653–1674.

Frasch, M., Chen, X., and Lufkin, T. 1995. Evolutionary-conserved enhancers direct region-specific expression of the murine Hoxa-1

- and Hoxa-2 loci in both mice and *Drosophila*. *Development* **121**: 957–974.
- Garcia-Fernandez, J. 2005. The genesis and evolution of homeobox gene clusters. *Nat. Rev. Genet.* **6**: 881–892.
- Gaudet, J., Muttumu, S., Horner, M., and Mango, S.E. 2004. Whole-genome analysis of temporal gene expression during foregut development. *PLoS Biol.* **2**: e352. doi: 10.1371/journal.pbio.0020352.
- Grant, K., Hanna-Rose, W., and Han, M. 2000. sem-4 promotes vulval cell-fate determination in *Caenorhabditis elegans* through regulation of *lin-39* Hox. *Dev. Biol.* **224**: 496–506.
- Graustein, A., Gaspar, J.M., Walters, J.R., and Palopoli, M.F. 2002. Levels of DNA polymorphism vary with mating system in the nematode genus *Caenorhabditis*. *Genetics* **161**: 99–107.
- Haerry, T.E. and Gehring, W.J. 1997. A conserved cluster of homeodomain binding sites in the mouse Hoxa-4 intron functions in *Drosophila* embryos as an enhancer that is directly regulated by Ultrabithorax. *Dev. Biol.* **186**: 1–15.
- Hillier, L.W., Coulson, A., Murray, J.I., Bao, Z., Sulston, J.E., and Waterston, R.H. 2005. Genomes in *C. elegans*: So many genes, such a little worm. *Genome Res.* **15**: 1651–1660.
- Hobert, O. 2002. PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic *C. elegans*. *Biotechniques* **32**: 728–730.
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241–254.
- Kelly, W.G., Xu, S., Montgomery, M.K., and Fire, A. 1997. Distinct requirements for somatic and germline expression of a generally expressed *Caenorhabditis elegans* gene. *Genetics* **146**: 227–238.
- Kimble, J. and Hirsh, D. 1979. The postembryonic cell lineages of the hermaphrodite and male gonads in *Caenorhabditis elegans*. *Dev. Biol.* **70**: 396–417.
- Kiontke, K., Hironaka, M., and Sudhaus, W. 2002. Description of *Caenorhabditis japonica* n. sp. (Nematoda: Rhabditida) associated with the burrowing bug *Parastrachia japonensis* (Heteroptera: Cydnidae) in Japan. *Nematology* **4**: 933–941.
- Kiontke, K., Gavin, N.P., Raynes, Y., Roehrig, C., Piano, F., and Fitch, D.H. 2004. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc. Natl. Acad. Sci.* **101**: 9003–9008.
- Kiontke, K., Barriere, A., Kolotuev, I., Podbilewicz, B., Sommer, R., Fitch, D.H., and Felix, M.A. 2007. Trends, stasis, and drift in the evolution of nematode vulva development. *Curr. Biol.* **17**: 1925–1937.
- Korf, I., Yandell, M., and Bedell, J. 2003. BLAST. O'Reilly, Sebastopol, CA.
- Krause, M., Harrison, S.W., Xu, S.Q., Chen, L., and Fire, A. 1994. Elements regulating cell- and stage-specific expression of the *C. elegans* MyoD family homolog *hlh-1*. *Dev. Biol.* **166**: 133–148.
- Krek, A., Grün, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., et al. 2005. Combinatorial microRNA target predictions. *Nat. Genet.* **37**: 495–500.
- Lemons, D. and McGinnis, W. 2006. Genomic evolution of Hox gene clusters. *Science* **313**: 1918–1922.
- Li, L., Stoeckert Jr., C.J., and Roos, D.S. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**: 2178–2189.
- Li, L., Zhu, Q., He, X., Sinha, S., and Halfon, M.S. 2007. Large-scale analysis of transcriptional *cis*-regulatory modules reveals both common features and distinct subclasses. *Genome Biol.* **8**: R101. doi: 10.1186/gb-2007-8-6-r101.
- Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. 2003. The microRNAs of *Caenorhabditis elegans*. *Genes & Dev.* **17**: 991–1008.
- Malicki, J., Cianetti, L.C., Peschle, C., and McGinnis, W. 1992. A human HOX4B regulatory element provides head-specific expression in *Drosophila* embryos. *Nature* **358**: 345–347.
- Maloof, J.N. and Kenyon, C. 1998. The Hox gene *lin-39* is required during *C. elegans* vulval induction to select the outcome of Ras signaling. *Development* **125**: 181–190.
- McGaughey, D.M., Vinton, R.M., Huynh, J., Al-Saif, A., Beer, M.A., and McCallion, A.S. 2008. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at *plox2b*. *Genome Res.* **18**: 252–260.
- McGhee, J.D., Sleumer, M.C., Bilenky, M., Wong, K., McKay, S.J., Goszczynski, B., Tian, H., Krich, N.D., Khattra, J., Holt, R.A., et al. 2007. The ELT-2 GATA-factor and the global regulation of transcription in the *C. elegans* intestine. *Dev. Biol.* **302**: 627–645.
- McKay, S.J., Johnsen, R., Khattra, J., Asano, J., Baillie, D.L., Chan, S., Dube, N., Fang, L., Goszczynski, B., Ha, E., et al. 2003. Gene expression profiling of cells, tissues, and developmental stages of the nematode *C. elegans*. *Cold Spring Harb. Symp. Quant. Biol.* **68**: 159–169.
- Mello, C. and Fire, A. 1995. DNA transformation. *Methods Cell Biol.* **48**: 451–482.
- Mortazavi, A., Leeper Thompson, E.C., Garcia, S.T., Myers, R.M., and Wold, B. 2006. Comparative genomics modeling of the NRSF/REST repressor network: From single conserved sites to genome-wide repertoire. *Genome Res.* **16**: 1208–1221.
- Okkema, P.G., Harrison, S.W., Plunger, V., Aryana, A., and Fire, A. 1993. Sequence requirements for myosin gene expression and regulation in *Caenorhabditis elegans*. *Genetics* **135**: 385–404.
- Olson, S. 2002. EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. *Bioinformatics* **3**: 87–91.
- Ovcharenko, I., Loots, G.G., Nobrega, M.A., Hardison, R.C., Miller, W., and Stubbs, L. 2005. Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15**: 137–145.
- Pauli, F., Liu, Y., Kim, Y.A., Chen, P.J., and Kim, S.K. 2006. Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in *C. elegans*. *Development* **133**: 287–295.
- Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**: 499–502.
- Popperl, H., Bienz, M., Studer, M., Chan, S.K., Aparicio, S., Brenner, S., Mann, R.S., and Krumlauf, R. 1995. Segmental expression of Hoxb-1 is controlled by a highly conserved autoregulatory loop dependent upon *exd/pbx*. *Cell* **81**: 1031–1042.
- Ruvinsky, I. and Ruvkun, G. 2003. Functional tests of enhancer conservation between distantly related species. *Development* **130**: 5133–5142.
- Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. 2004. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**: D91–D94.
- Sinha, S., Schroeder, M.D., Unnerstall, U., Gaul, U., and Siggia, E.D. 2004. Cross-species comparison significantly improves genome-wide prediction of *cis*-regulatory modules in *Drosophila*. *BMC Bioinformatics* **5**: 129. doi: 10.1186/1471-2105-5-129.
- Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigan, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12**: 1611–1618.
- Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., Roy, S., Deoras, A.N., et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**: 219–232.
- Sternberg, P.W. 2005. Vulval development. In *WormBook* (eds. The *C. elegans* Research Community). doi: 10.1895/wormbook.1.6.1, <http://www.wormbook.org>.
- Stone, E.A., Cooper, G.M., and Sidow, A. 2005. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **6**: 143–164.
- Stothard, P. and Pilgrim, D. 2006. Conspecific and interspecific interactions between the FEM-2 and the FEM-3 sex-determining proteins despite rapid sequence divergence. *J. Mol. Evol.* **62**: 281–291.
- Stoyanov, C.N., Fleischmann, M., Suzuki, Y., Tapparel, N., Gautron, F., Streit, A., Wood, W.B., and Muller, F. 2003. Expression of the *C. elegans* labial orthologue *ceh-13* during male tail morphogenesis. *Dev. Biol.* **259**: 137–149.
- Streit, A., Kohler, R., Marty, T., Belfiore, M., Takacs-Vellai, K., Vigano, M.A., Schnabel, R., Affolter, M., and Muller, F. 2002. Conserved regulation of the *Caenorhabditis elegans* labial/Hox1 gene *ceh-13*. *Dev. Biol.* **242**: 96–108.
- Sudhaus, W. and Kiontke, K. 1996. Phylogeny of Rhabditis subgenus *Caenorhabditis* (Rhabditidae, Nematoda). *J. Zoo. Syst. Evol.* **34**: 217–233.
- Sudhaus, W. and Kiontke, K. 2007. Comparison of the cryptic nematode species *Caenorhabditis brenneri* sp. n. and *C. remanei* (Nematoda: Rhabditidae) with the stem species pattern of the *Caenorhabditis elegans* group. *Zootaxa* **1456**: 45–62.
- Sulston, J. and Hodgkin, J. 1988. Methods. In *The nematode Caenorhabditis elegans* (ed. W.B. Wood), pp. 587–606. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
- Sulston, J.E. and Horvitz, H.R. 1977. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* **56**: 110–156.
- Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., and Jones, R.T. 1988. Embryonic and globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203**: 439–455.
- Tomba, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E.,

- Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23**: 137–144.
- Wagmaister, J.A., Miley, G.R., Morris, C.A., Gleason, J.E., Miller, L.M., Kornfeld, K., and Eisenmann, D.M. 2006. Identification of *cis*-regulatory elements from the *C. elegans* Hox gene *lin-39* required for embryonic expression and for regulation by the transcription factors LIN-1, LIN-31 and LIN-39. *Dev. Biol.* **297**: 550–565.
- Wang, X. and Chamberlin, H.M. 2004. Evolutionary innovation of the excretory system in *Caenorhabditis elegans*. *Nat. Genet.* **36**: 231–232.
- Wang, T. and Stormo, G.D. 2003. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19**: 2369–2380.
- Wang, B.B., Muller-Immergluck, M.M., Austin, J., Robinson, N.T., Chisholm, A., and Kenyon, C. 1993. A homeotic gene cluster patterns the anteroposterior body axis of *C. elegans*. *Cell* **74**: 29–42.
- Wei, C., Lamesch, P., Arumugam, M., Rosenberg, J., Hu, P., Vidal, M., and Brent, M.R. 2005. Closing in on the *C. elegans* ORFeome by cloning TWINSCAN predictions. *GenomeRes.* **15**: 577–582.
- Wenick, A.S. and Hobert, O. 2004. Genomic *cis*-regulatory architecture and *trans*-acting regulators of a single interneuron-specific gene battery in *C. elegans*. *Dev. Cell* **6**: 757–770.
- Wittmann, C., Bossinger, O., Goldstein, B., Fleischmann, M., Kohler, R., Brunschwig, K., Tobler, H., and Muller, F. 1997. The expression of the *C. elegans* labial-like Hox gene *ceh-13* during early embryogenesis relies on cell fate and on anteroposterior cell polarity. *Development* **124**: 4193–4200.
- Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**: e7. doi: 10.1371/journal.pbio.0030007.
- Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**: 338–345.
- Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., Kellis, M., and Lander, E.S. 2007. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl. Acad. Sci.* **104**: 7145–7150.
- Zhao, G., Schriefer, L.A., and Stormo, G.D. 2007. Identification of muscle-specific regulatory modules in *Caenorhabditis elegans*. *Genome Res.* **17**: 348–357.

Received August 26, 2008; accepted in revised form September 17, 2008.

**Chapter 3: Synthetic PAT screen reveals additional
myogenic transcription factors**

Synthetic PAT screen reveals additional myogenic transcription factors

Steven Kuntz^{1,2}, Emily Riggall^{1,3}, Barbara Wold¹, Paul Sternberg^{1,2}

1 Division of Biology and 2 Howard Hughes Medical Institute, California Institute of Technology, Pasadena, CA 91125, U.S.A.

3 Current Location: Carlton College, Northfield, MN 55407, U.S.A.

ABSTRACT

To help identify novel components of the body wall muscle differentiation transcription factor network in *C. elegans*, we used an RNAi screen for synthetic lethality. Genetic and molecular studies, primarily focused on individual components, have revealed several muscle differentiation factors, including *hlh-1*, *unc-120*, and *hnd-1*. However, given the stability of the network in the presence of mutations, additional components of the network are best revealed only when the network is compromised. We conducted a synthetic lethal screen, using an RNAi library to knock down genes encoding transcription factors in an *hlh-1* mutant background. This screen identified several transcription factors that were likely to function in muscle differentiation. We then analyzed the positives with the strongest phenotypes using existing GFP expression, *in situ* hybridization, and microarray expression data in order to determine their putative interactions with other components of the differentiation network. Transcription factors such as *hmg-1.2*, *tbx-33*, *fkh-3*, *tbp-1*, *nhr-46*, *nhr-71*, *nhr-112*, D1046.2, *nhr-134*, Y6G8.3, and F52C12.4 were identified that may play a role in muscle, but whose specific function within muscle is unknown. Other transcription factors are known to play at least

some role in muscle development, including *ceh-20*, *ceh-49*, *ceh-51*, *grh-1*, and *lin-1*. Other factors, though exhibiting a synthetic lethal effect, have unclear roles in muscle: *sex-1*, *cnd-1*, and *sdc-2*.

INTRODUCTION

Biological networks are sometimes remarkably resistant to injury, being able to accomplish their task even when parts of the network are removed. To understand how different parts of the system interact to compensate for such aberrations requires a thorough understanding of the network's foundation. The different parts of the system must first be identified. The identification and description of additional components will help us construct a draft map of the muscle differentiation network and facilitate a more complete understanding of the network architecture.

The muscle differentiation network is composed of a number of known and unknown transcription factors. Transcription factors within this network cooperate in such a way that single mutations do not completely collapse the network, giving the network a rugged character. However, to what extent such rugged behaviour is due to overlapping transcription factor functions or other compensatory regulatory effects is unknown. We decided to focus on the network's transcription factors to further expand our knowledge of what genes are involved. Because no single mutation will stop all muscle differentiation, single mutant screens are of limited use in identifying necessary factors (Baugh et al., 2005b; Fukushige et al., 2006). To identify these other factors, it is

necessary to sufficiently compromise the network through mutation or RNAi such that an additional mutation would the network to collapse.

Our current understanding of transcription factors acting in the differentiation of nematode body wall muscle is highly informed by homologs in other phyla. Operating under the assumption that numerous genes are conserved between all muscle, the major nematode components were gradually identified (Harfe et al., 1998a; Harfe and Fire, 1998; Harfe et al., 1998b). Though the roles of the factors vary a little between phyla, the same families of proteins appear present (Figure 1). For instance, an Nkx-2.5/Tinman protein controls mammalian and fly cardiac differentiation as well as nematode pharyngeal differentiation. Genes involved in skeletal and smooth muscle fall into the same categories as the body wall and non-striated muscles of *C. elegans*: MADS, MRF, and Twist. The gene *hnd-1* in *C. elegans* present a special case, as it also regulates a non-muscle tissue (gonad development) (Mathies et al., 2003). *hnd-1* functions earlier in the process of differentiation than the other genes and may play a different role (Fukushige et al., 2006). The orthologs for *hnd-1*, the genes HAND1 and HAND2 in vertebrates play a similar specification role, but not in skeletal muscle.

Synthetic interactions are emergent phenotypes that are only observed with the combined impact of multiple mutations as a result of the mutations affecting overlapping or compensatory pathways. Either single mutation often has a minor effect. However, most animals will not exhibit the phenotype of interest. When the two mutations are introduced into the same animal the phenotype is amplified. This is thought to occur because there are overlapping pathways controlling a phenotype such that if either pathway fails, the other will take over. However, mutations in both pathways cause both

to collapse, meaning no pathways can maintain the wild type phenotype and the synthetic phenotype will emerge. If two pathways are completely independent and non-overlapping, the penetrance of the phenotype should be additive. If they are within the same pathway, it is expected that there would be no increase or silencing in the phenotype. However, if they are in overlapping pathways the penetrance of the phenotypes should exceed the fractional product of the two component phenotypes (Figure 2).

In determining whether mutations interact, we start with the null hypothesis (Baugh et al., 2005b), that there is no interaction between the two networks:

$$(1 - \text{Fractional phenotype in mutant}) * (1 - \text{Fractional phenotype from RNAi in wildtype}) = (1 - \text{Fractional phenotype from RNAi in wildtype})$$

This can then be modified to generate a scoring system based on the observations:

$$\text{Score} = (1 - \text{Fractional phenotype in mutant}) * (1 - \text{Fractional phenotype from RNAi in wildtype}) / (1 - \text{Fractional phenotype from RNAi in wildtype})$$

With this set up, the value of the score is easily parsed into three categories:

If score < 1, then there is phenotype suppression or genes are in the same pathway

If score > 1, then there is a synthetic lethal interaction

If score = 1, then there is no overlap of the pathways.

We are interested in mutations that are relevant to muscle formation, specifically muscle differentiation. During the development of *C. elegans*, embryos start out as

ellipsoid and depend on surface constriction to elongate and form a full-length larval worm (Figure 3A). Initially the development and closing of the epidermis constricts the embryo, driving the embryo from a bean-shape to a more horseshoe shape, termed the two-fold stage. At this point the epidermis can no longer constrict the worm and the differentiation of muscle is required to continue elongation to a longer worm, termed the pretzel stage, named for its looped shape. If the muscle does not terminally differentiate and form contractile tissue, the animal will not properly elongate. Incomplete elongation leads to dumpy worms (worms that are shorter and squatter than wild type), and if the muscle is malformed, the animals typically form what are known as lumpy-dumpy animals (Figure 3B). Lumpy dumpy animals are shorter than wild type animals and are uneven in their elongation, giving a lumpy appearance. They do hatch and can make small movements or twitch. Some severe muscle differentiation mutations, such as *hlh-1(cc561)*, give this phenotype, as they can cause disorder in the differentiation but do not stop it completely in most cases. We are looking for a phenotype that indicates that all muscle differentiation has stopped: a complete failure in muscle-dependent elongation. We see this in small numbers, but are screening for such a phenotype in large numbers at high penetrance. This phenotype is characterized by a complete paralysis at the two-fold stage, the PAT phenotype (Figure 3B). Such animals are completely unable to move except for the contractions of the pharynx. Pharyngeal muscle regulation is independent of body wall muscle development. This is not an immediately lethal phenotype and the animals continue to develop and grow, at times even hatching (which is an enzymatic process independent of movement). Nevertheless they remain immobile and horseshoe-shaped.

In this study we investigated the effects of synthetic lethal mutations with *hlh-1* and identified a number of genes that appear to act in conjunction with or in place of *hlh-1* in body wall muscle differentiation.

RESULTS

To identify genes that interact with *hlh-1* during muscle differentiation, we performed an RNAi screen against transcription factors in the Ahringer lab RNAi feeding library (Fraser et al., 2000). At the time that the screen was performed, there were 513 transcription factors in the RNAi library out of 934 transcription factors (Figure 4) (Reece-Hoyes et al., 2005). The estimate is based on transcription factor predictions and is highly dependent on DNA-binding motif predictions. The screen was by no means comprehensive, but we expected it would still give significant insight into body wall muscle differentiation given the coverage of factors (Fernandes and Sternberg, 2007).

For the screen we compared the level of lethality in wild type animals fed the RNAi with the level of lethality (and specifically a PAT phenotype). Because the *hlh-1* mutants have a certain baseline level of PAT embryos, for a positive hit it was necessary to have a significantly higher level of PAT phenotype in the mutant when fed the RNAi.

For a number of genes, the RNAi proved to be lethal even in the wild-type animals (Table 1). This lethality was typically in early embryonic development, prior to the two-fold stage. Therefore many of these genes are necessary for early development that precedes muscle specification and differentiation. These genes present a problem in that they may be active in muscle differentiation, but it is impossible to tell with this

assay. Several well-known early embryonic determinants of specification are included in this list, such as *pal-1*, *mex-3*, and *pie-1*. As expected, there were no genes whose RNAi produced a PAT phenotype in the wild type animals.

Most genes exhibited no phenotype. This does not exempt them from involvement in muscle differentiation. As RNAi assays have a very high false-negative rate, it is possible that many of these genes were not sufficiently knocked down to cause a phenotype. Due to the general viability of the *hll-1* mutants, no suppressors of the *hll-1* mutation were found either, as the statistical requirements were too strict. The only genes expected to interact with *hll-1* gave a synthetic interaction: *unc-120* and *hnd-1*. Neither *hll-8* nor *ceh-22* had a synthetic interaction. This was expected since they are the major myogenic factors only within non-body wall muscle and their regulation is not expected to overlap.

The screen identified a number of genes that had an increased level of lethality in the mutant when fed RNAi, indicating that they are necessary for buffering the differentiation of muscle in the nematodes. Of the 513 genes screened, 40 genes were selected (Table 2) as being the best new candidates for exhibiting a synthetic lethal interaction with *hll-1*.

The coverage of the OpenBioSystems library was slightly different from that of the Ahringer library. It is an independently created library and may have its own biases. We screened 78 genes from the library and identified 27 genes for further analysis, including 7 genes overlapping with the Ahringer set.

The top candidates were re-screened, along with controls. For the second round of screening we paid close attention to the nature of the phenotype. Whereas the initial screen was largely looking for synthetic lethality, this screen was investigating the PAT phenotype. With the new screen, the number of candidate genes was reduced to 44 total candidate genes.

Some genes were screened against *hnd-1(q740)* mutants as well. HND-1 is expressed earlier in embryonic development, at the early stages of muscle differentiation (Table 3). Since *hnd-1* has some overlapping properties with *hlh-1* (including the fact that HND-1 is believed to bind to early HLH-1 targets prior to the activation of *hlh-1* (Fukushige, 2006 #18)), it is expected that some of the same genes appearing in the synthetic lethal screen against *hlh-1* mutants will also demonstrate a synthetic phenotype against the *hnd-1* mutants. However, the overlap is not expected to be complete, as the genes act in slightly different ways and any genes that act with *hnd-1* will show up only as acting synthetically with *hlh-1*. The screen turned up four genes as interacting with *hnd-1*: *exc-9*, *nhr-4*, *lin-14*, and *tbp-1*. These overlap somewhat with *hlh-1* hits, but except for *tbp-1* they are not the top hits of the screen.

A smaller set of genes was also screened against *unc-120(st364)* mutants. UNC-120 acts with *hlh-1* to be a major myogenic factor in differentiation (Baugh et al., 2005b; Williams and Waterston, 1994). It is expressed in all non-pharyngeal muscle, so it may have different targets from HLH-1 and therefore interact with a different set of genes. From the limited screen only four genes showed an interaction with *unc-120*: *ceh-49*, *ceh-51*, *tbp-1*, and *nhr-134*. All of these genes also showed interactions with *hlh-1* and are described below.

To identify which of these genes would be the best targets for future study we compared the existing data regarding the gene candidates. Available data consists of microarray expression data from embryonic muscle (Baugh et al., 2005a), larval muscle (Roy et al., 2002), and whole embryos (Hill et al., 2000; Kim et al., 2001); GFP (green fluorescent protein marker driven by gene promoters) expression data; and *in situ* hybridization data (Tabara et al., 1996). Any of the genes with expression observed in the body wall muscle of the embryo was flagged. Some of the expression data was not very precise, sometime remarked as simply the late embryo or as body wall muscle. For the sake of completeness, such observations were included. GFP expression was given the highest significance regarding expression levels due to its greater precision both spatially and temporally. *In situ* hybridization, though accurate, is not as precise in the worm. Finally, the microarray data was given the lowest significance due to the lack of specificity in the samples.

In total, 20 of the candidate genes were observed to have expression in the correct time or place (Figure 5). 10 of these genes had GFP expression in the correct location, making them slightly higher priority. Of those, 5 also had *in situ* expression. And of those, only 3 also had the correct microarray expression. These hits were then ranked according to their success in the screen.

Four of the candidate genes – *ceh-51*, *hmg-1.2*, *sex-1*, and *ceh-20* – were screened with dsRNA injections to further investigate their interactions with different members of the myogenic network. This technique in some cases gives a stronger phenotype than RNAi feeding (Hunter, 1999), ideal for looking at small numbers of worms.

The most promising hit by the ranking criteria was the gene *hmg-1.2*, also known as *son-1*. It has previously been described for its role in Wnt signalling (Jiang and Sternberg, 1999). This may relate to its more recently described role as a “hub” protein (Lehner et al., 2006). As a hub protein, it interacts with numerous networks, one of which is Wnt signalling. It is possible that the Wnt signalling is important for a pathway involved in muscle differentiation. It might also help in regulating pathways that serve to compensate muscle differentiation in the absence of *hlh-1*. When *hmg-1.2* dsRNA was injected in different mutant backgrounds it showed a strong synthetic PAT interaction with the *hlh-1(cc561)* mutant but no synthetic PAT interaction with *unc-120(st364)*. This may indicate that the compensation pathway for an *hlh-1* mutation differs significantly from the compensation pathway in *unc-120* mutants. Another possibility is that *unc-120* and *hmg-1.2* are in the same pathway, meaning that knocking down both in conjunction will give no additive effect.

The second most promising hit was *ceh-20*. This gene is an ortholog of Extradenticle in the Pairedbox family of transcription factors. It is known to interact with *unc-62* and *ceh-40* to work with Hox genes, such as *lin-39* and *mab-5* (Jiang et al., 2009; Potts et al., 2009). Both of these genes are involved in muscle differentiation. This makes *ceh-20* a very strong candidate for muscle regulation. *ceh-20* has previously been described as controlling the fate decision of neuron cell death (Liu et al., 2006) and therefore may also play a role in muscle fate decision when the network has been compromised. A strain of *ceh-20* mutants was acquired and injected with dsRNA from several different transcription factors. A weak synthetic PAT interaction was seen with *hlh-1*, *unc-120*, and *hmg-1.2* RNAi, but the significance was limited. This may be

reflective of the potency of the mutation. A stronger mutation may give a stronger phenotype. However, due to the broad roles of *ceh-20* in early development other lethality issues may mask any synthetic lethality.

The third candidate was *ceh-49*. It is a member of the onecut homeobox genes. Little is known about it, though it is strongly expressed in much of the early embryo. Expression fades slowly as the animal ages, but it still had significant expression at hatching (Liu et al., 2009; Reece-Hoyes et al., 2007).

The fourth candidate was *ceh-51*, also previously known as *dlx-1* and Y80D3A.3. The gene has been recently described as an important part of the muscle regulatory network, especially in the MS lineage, which gives rise to roughly one third of the worm's embryonically derived muscle. Another group was simultaneously studying its role in specification (please see the Appendix, (Broitman-Maduro et al., 2009)). Though there is no convincing microarray data, there is strong GFP and *in situ* data to support its placement in the correct muscle precursors. To analyze its interaction with other transcription factors, we injected dsRNA for the major myogenic factors into the *ceh-51(tm2123)* mutant strain (Figure 6). We also injected the *ceh-51* dsRNA into different mutant backgrounds (Figure 6). Since *ceh-51* is active in the MS muscle progenitors (Broitman-Maduro et al., 2009), it has a very limited synthetic PAT interaction with *hnd-1*, which is active primarily in the C and D lineage muscle progenitors (Baugh et al., 2005b). Therefore this gene may serve as a counterpart to *hnd-1*. The synthetic lethal interactions with *unc-120* and *hlh-1* were relatively strong (Figure 7). This indicates that it, unlike *hmg-1.2*, is in an independent pathway from both *unc-120* and *hlh-1*. However,

the synthetic phenotype demonstrates that it cannot compensate for both factors missing. This data reinforces the proposal that this is an *hnd-1* counterpart.

The fifth candidate was *sex-1*. Its role in interacting with muscle progenitors is unclear, as it is primarily involved in sex determination. *Sex-1* is a nuclear hormone receptor that acts to repress *xol-1*, leading to development as a hermaphrodite. It is expressed in all nuclei from oogenesis through the mid-embryo, which is when the body wall muscle is formed. Therefore it is present in the muscle progenitors and may interact with muscle differentiation factors or its control of sex determination may be necessary for other compensatory transcription factors to function in *hll-1* mutants. A *sex-1* mutant was injected with dsRNA to observe the synthetic PAT phenotype prevalence. However, very few animals exhibited the PAT phenotype. Again, this may relate to the nature of the mutation.

The remaining candidates were not analyzed in as great of detail. The gene *lin-1* is important for MAP kinase signal transduction. It is expressed in the body wall muscle, but has primarily been studied in the vulva, where it acts to repress fate specification of vulval cells. It may play a similar role in muscle development, perhaps being involved in repressing inhibitors of muscle differentiation.

Interestingly, *tbp-1* gave a consistently strong synthetic PAT phenotype despite it coding for the TATA binding protein. It is expected to have a broad impact on transcription, especially in the embryo. The specific effect on muscle is unexplained, but may relate either to the special character of muscle requiring large amounts of transcription or may lead to sickly cells simply ceasing function in its absence.

cnd-1 is the ortholog of mammalian NeuroD. It is a helix-loop-helix protein just like *hlh-1*. In fact, it binds to virtually the same motif (Grove et al., 2009). It is possible that in the absence of *hlh-1*, *cnd-1* becomes important in activating transcriptional targets since they should bind to virtually the same sequences. What the mechanism of this or in what way such compensation would develop is unknown. Some muscle-specific expression of *cnd-1* has been observed (Liu et al., 2009), meaning the possibility of it playing a role is plausible.

Another regulator of cell fates, *grh-1* is also a candidate. *grh-1* encodes a Grainyhead gene and regulates such Hox genes as *mab-5* (Venkatesan et al., 2003). As the Hox genes are intricately involved in tissue development and fate specification, *grh-1* may play a significant role in this network.

The gene *tbx-33* is not well described, but appears to have its expression relatively restricted to the body wall muscle (Liu et al., 2009). It also appears to have a binding site for MED-1, a regulatory GATA factor, which may play a role in its expression in the EMS lineage (Broitman-Maduro et al., 2005). Many T-box factors are important for fate regulation both in *C. elegans* and in other species.

The genes *fkh-3*, *nhr-46*, *nhr-71*, F52C12.4, *nhr-112*, D1046.2, *nhr-134*, and Y6G8.3 are not well described and their functions remain unknown. Future studies may reveal more about these factors.

DISCUSSION

We have identified a number of genes that have varying degrees of importance in muscle development. At least one of the genes, *ceh-51*, has proven to be an important specification and early differentiation factor in embryonic muscle development (see Appendix (Broitman-Maduro et al., 2009)). Other genes are known to play a role in muscle development or serve as general regulatory factors, so their inclusion in the muscle set is not surprising. Other genes identified, however, are very surprising because they have never been characterized in the regulation of muscle formation.

One major advantage of using RNAi to perform such a screen in nematodes is the ability to more rapidly screen through factors than could be accomplished in either vertebrates or insects. RNAi is far from perfect; it has a very high false negative rate. There are several steps in the RNAi feeding process that may result in a failure of the RNAi to properly knock down the target gene. Uptake of the RNAi is a major issue. Properly targeting the intended gene can vary based on the sequence composition or on the splicing variants for the gene. Additionally, certain tissues do not respond as well to RNAi as other tissues. For instance, the neurons of *C. elegans* are very difficult targets for RNAi knock down while the muscle and intestine are very easy targets. Thus, using RNAi in a muscle differentiation screen should have fewer difficulties than in other tissues. A striking advantage of RNAi is the extremely low false positive rate. If a result is seen, it is most likely a real result rather than an artefact (Fraser et al., 2000).

What may prove the most novel is the appearance of the sex determination and dosage compensation genes *sex-1* and *sdc-2*. These genes play a major role in regulation of transcription by silencing large sections of the X chromosome (Nusbaum and Meyer, 1989). Their appearance coupled with the screen identifying *tbp-1* may suggest that

damage to basic transcriptional machinery can present a problem in *hlh-1* mutants. The mutant transcriptional system may already be stressed and the remaining differentiation pathway may not be capable of handling any defects in the transcriptional machinery. Nevertheless, it is unlikely that all of the candidate proteins simply make a sick animal a little bit sicker. Many of the genes fit into distinct regulatory categories that have little to do with the process of transcription, but rather have more to do with its regulation.

The sex determination/dosage compensation proteins and neurogenic factor *cnd-1* may be present as part of a transcriptional network that is co-opted in muscle specification. The expression of *cnd-1* in muscle tissue has not been explained mechanistically, so it is possible that these proteins have further roles than those known and described. CND-1 does have the property of binding to the same E-box *in vitro* as HLH-1 (Grove et al., 2009). Therefore its activation may serve to compensate for the loss of HLH-1 binding. The mechanism by which these different network architectures are activated is not known. One possibility is that master regulators involved in general patterning, such as the Hox genes and their cofactors, are involved in such an activation process. A Hox cofactor, *ceh-20*, and Hox activator, *grh-1*, are known to be active in the muscle (Jiang et al., 2009; Potts et al., 2009; Venkatesan et al., 2003) and are identified by the screen. They may serve to recruit genes not normally known to function in muscle differentiation (Figure 8). These Hox-associated factors are known to act in molecular switches. If a major myogenic factor is missing, it is understandable that differentiation might best continue if the expression of major transcription factors is switched. This switch may be flipped through either expression of a gene normally repressed by *hlh-1* or by activation of a gene that normally depends on feedback from *hlh-1*.

The cross network activation may explain the presence of *hmg-1.2*. Being widespread and involved as a “hub” in multiple networks suggests a role for *hmg-1.2* in the synthetic PAT phenotype. Rather than performing any vital duties for specifically muscle transcription, it may play the role of a middleman. *hmg-1.2* may simply be necessary for the compensation pathway to properly remedy the network. Alternatively, it may exist as a hub protein due to a very important role in differentiation in multiple tissues that is necessary in compromised animals.

Overall, several good candidate additions to the muscle transcriptional network were suggested by the results of the screen and at least one new transcription factor, *ceh-51*, was added. With this study we can expand the field of transcription factors known to function in muscle specification and differentiation (Figure 9). None of the additional factors are exclusively functional in muscle, but their role is clearly critical.

The homologs of the muscle transcription factors are likely to follow in suit of *hnd-1* rather than *hlh-1* and *unc-120*. Because they are not muscle-specific and are involved more in specification than terminal differentiation, both *hnd-1* and *ceh-51* will likely have orthologs in vertebrates and insects that perform similar roles in fate determination, but not necessarily in the same tissue. This can be seen with *hnd-1*, which has a dual role in muscle formation and gonadogenesis (Mathies et al., 2003). Its homologs in vertebrates, HAND1 and HAND2, are involved in fate specification, but in the heart and other tissues rather than in the skeletal muscle. The closest orthologs for *ceh-51* are the vertebrate genes *Dlx-1*, *Barx1*, and *Bsx* and the insect gene *CG7056*. These genes are not known to function in muscle differentiation, but they may play important roles in tissue specification and would be worthwhile targets for further study.

Similarly, *hmg-1.2* is not specific to muscle and its orthologs may also serve as hub proteins that are important in multiple networks. Its orthologs include vertebrate HMGB and insect Dsp1. *lin-1*, already known for its role in fate determination for vulval cells, may prove important for a similar role in other organisms.

MATERIALS AND METHODS

General methods and strains. We obtained *Caenorhabditis elegans* from the CGC strain collection and cultured them on OP50 at 20°C, using methods standard for *C. elegans* (Sulston and Hodgkin 1988). MS1208 worms were generated by microinjection of *unc-119::mCherry*, *myo-2::mCherry*, *ceh-51*, and carrier. MS1208 hermaphrodites were subsequently microinjected with a mixture of 25 ng/μL unpurified PCR amplified Fire Lab Vector pPD93.48 (*unc-54::gfp*) and 150 ng/μL pBluescript to generate transgenic animals (Kelly et al. 1997; Mello and Fire 1995).

RNAi feeding. Bacteria from the OpenBioSystems RNAi library and the Ahringer RNAi library were used for RNAi feeding of L4 animals for 36 hours at 25°C. Adults were then transferred to fresh plates for egg-laying for 4 hours at 25°C. Adults were removed and embryos were allowed to develop for 18-24 hours prior to scoring.

dsRNA injection. Standard T7 primers were used to amplify *mex-3*, *skn-1*, *pie-1*, *hlh-1*, *hnd-1*, *unc-120*, and *ceh-51*(Y80D3A.3) from the Ahringer Lab RNAi Library (Kamath and Ahringer 2003) using the Roche Expand High Fidelity PCR system. *mex-1* was amplified with custom primers from genomic DNA (TAATACGACTCACTATAGGAGCGAGTACAACCGTGCTCT,

TAATACGACTCACTATAGGCGGACTAACTGGTTTTCCGA) using the Roche Expand High Fidelity PCR system. The Ambion MEGAscript T7 High Yield Transcription kit was used to generate dsRNA. dsRNA was then microinjected into late L4 worms, as described in wormbook (Ahringer 2006). Injected animals were kept at 25°C for 3 hours. They were then transferred to fresh plates for egg-laying at 25°C for 21 hours, at which time the injected animals were removed and bleached to isolate their eggs. Eggs from the bleaching and the 21-hour egg-lay were pooled for analysis. Embryos were allowed to develop for 18-24 hours prior to scoring.

Scoring. Embryos were scored for developmental progression using a dissecting microscope. The stage of developmental arrest in embryonic lethal worms was noted as during the two-fold stage (PAT) or otherwise. MS1208 animals were screened for mCherry fluorescence to guarantee that only embryos not carrying the rescue construct were scored.

Nomarski imaging. Transgenic animals were viewed with Nomarski optics and a Chroma High Q EnGFP LP, FITC, or Texas Red filter cube on a Zeiss Axioplan, with a 100X oil objective, an X-cite series 120 UV epifluorescence light source, and a Hamamatsu ORCA II digital camera using Improvise Openlab software. ImageJ v1.37 was used to adjust image brightness and contrast and generate overlays. MS1208 embryos were freeze-cracked on dry ice and fixed in 4% formaldehyde, then stained with phalloidin-alexafleur 488. MS1208 embryos with *unc-54::gfp* and stained embryos were both imaged on 2% noble agar.

ACKNOWLEDGEMENTS

Some nematode strains used in this work were provided by the Caenorhabditis Genetics Center, which is funded by the NIH National Center for Research Resources (NCRR). We would like to thank LR Baugh for discussion on experimental setup and M Maduro for the MS1208 strain as well as consultation on the experimental setup.

TABLES

Table 1: Genes with high lethality in wild type animals

	Eggs	Larvae	Percent Lethal
mex-3	63	0	100.0
taf-5	340	0	100.0
cdk-9	122	1	99.2
lin-26	40	0	100.0
icd-1	129	20	86.6
bra-2	174	0	100.0
arx-6	80	0	100.0
pal-1	345	0	100.0
R07E5.3	249	0	100.0
cbp-1	67	0	100.0
T16H12.4	22	3	88.0
taf-9	17	8	68.0
pie-1	47	0	100.0
elc-1	70	0	100.0
W02C12.3	35	27	56.5
skn-1	24	0	100.0

lag-1	38	8	82.6
B0496.7	52	16	76.5
spt-5	12	0	100.0
mex-5	11	0	100.0
nhr-127	35	60	36.8
pha-4	73	41	64.0
unc-62	55	2	96.5
taf-10	16	6	72.7
pos-1	3	0	100.0
ZK1193.5	13	50	20.6

Table 2: Genes showing a synthetic lethal and PAT interactions

This table shows the values for synthetic lethal scoring in *hlh-1(cc561)* mutants. The score is determined by $(1 - \text{Fractional PAT in mutant}) * (1 - \text{Fractional PAT from RNAi in wildtype}) / (1 - \text{Fractional PAT from RNAi in wildtype})$. Genes with a score above 1.1 were determined to be significant and are shown in bold. Genes below the threshold are italicized.

RNAi	Wild type (N2)			<i>hlh-1(cc561)</i>			Score
	Lethal	PAT	Elongated	Lethal	PAT	Elongated	
<i>C0H6.5</i>	<i>10</i>	<i>13</i>	<i>121</i>	<i>10</i>	<i>5</i>	<i>57</i>	<i>0.929</i>
cnd-1	0	4	45	9	7	13	1.150
D1086.2	0	0	48	20	30	55	1.330
exc-9	6	0	112	2	33	3	7.220
ceh-49	0	2	91	17	8	17	1.148

F28C6.2	0	0	93	8	20	8	2.138
<i>nhr-4</i>	17	2	202	7	4	36	1.029
F3304.1	3	0	70	5	7	30	1.140
<i>F5401.4</i>	11	2	77	1	5	35	1.058
hlh-19	0	0	12	3	7	12	1.393
hmg-1.2	10	2	331	18	58	237	1.159
hnd-1	1	0	125	9	16	63	1.161
hnd-1	0	2	76	3	12	20	1.409
<i>lin-14</i>	4	2	110	14	1	111	0.941
<i>lin-26</i>	35	2	1	21	1	0	0.943
<i>mex-3</i>	135	0	0	121	0	0	0.950
nhr-11	3	0	82	0	9	16	1.484
nhr-134	4	0	65	8	29	44	1.480
nhr-46	0	0	136	5	14	13	1.689
nhr-60	0	7	135	0	12	27	1.305
nhr-63	3	2	136	4	2	4	1.171
<i>oma-2</i>	12	3	116	12	4	13	1.077
sex-1	1	1	32	32	21	9	1.394
T03E6.3	17	0	30	7	22	37	1.425
T5F2A.4	0	0	79	2	8	8	1.710
taf-5	0	1	145	12	11	39	1.147
tbp-1	0	0	32	22	6	4	1.169
unc-120	12	0	577	0	3	16	1.128
Y62E10a.17	0	0	224	12	17	2	2.104
ceh-51	0	0	132	16	12	32	1.188

Table 3: Genes showing a synthetic PAT interaction in *hnd-1* mutants

This table shows the values for synthetic lethal scoring in *hnd-1(q740)* mutants. The score is determined by (1-Fractional PAT in mutant) * (1-Fractional PAT from RNAi in wildtype) / (1-Fractional PAT from RNAi in wildtype). Genes with a score above 1.1 were determined to be significant and are shown in bold. Genes below the threshold are italicized.

RNAi	Wild type (N2)			<i>hnd-1(q740)</i>			Scores
	Lethal	PAT	Elongated	Lethal	PAT	Elongated	
<i>cnd-1</i>	0	4	45	3	0	31	0.891
<i>D1086.2</i>	0		48	11	7	120	1.022
exc-9	6	0	112	10	18	76	1.173
<i>F28C6.2</i>	0	0	93	12	2	40	1.007
nhr-4	17	2	202	13	19	42	1.293
<i>F3304.1</i>	3	0	70	9	3	46	1.023
<i>hlh-1</i>	0	0	107	13	5	54	1.042
<i>hmg-1.2</i>	7	0	110	5	1	50	0.988
<i>hnd-1</i>	0	2	76	12	1	45	0.962
lin-14	4	2	110	6	33	39	1.652
<i>lin-26</i>	35	2	1	7	1	67	0.931
<i>nhr-11</i>	3	0	82	1	0	30	0.970
<i>nhr-134</i>	4	0	65	2	0	5	0.970
<i>nhr-46</i>	0	0	136	8	0	1	0.970
<i>oma-2</i>	12	3	116	5	0	17	0.948
<i>sex-1</i>	1	1	32	7	5	30	1.069
<i>T03E6.3</i>	17	0	30	5	0	28	0.970
<i>T5F2A.4</i>	0	0	79	9	0	20	0.970

<i>taf-5</i>	0	1	145	3	2	41	1.007
tbp-1	0	0	32	0	5	18	1.239
<i>Y62E10a.17</i>	0	0	224	2	4	54	1.039
<i>ceh-51</i>	0	0	132	4	0	31	0.970

Table 4: Genes showing a synthetic PAT interaction in *unc-120* mutants

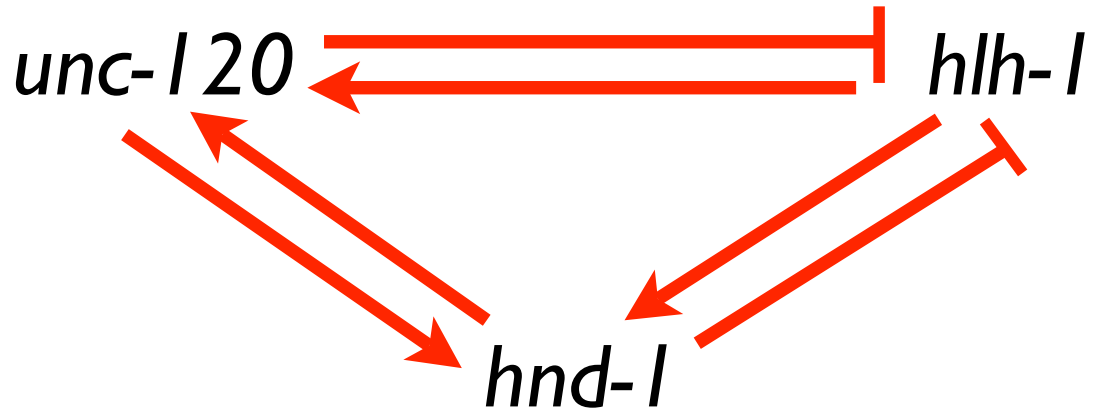
This table shows the values for synthetic lethal scoring in *unc-120(st364)* mutants. The score is determined by (1-Fractional PAT in mutant) * (1-Fractional PAT from RNAi in wildtype) / (1-Fractional PAT from RNAi in wildtype). Genes with a score above 1.1 were determined to be significant and are shown in bold. Genes below the threshold are italicized.

RNAi	Wild type (N2)			<i>unc-120(st364)</i>			Scores
	Lethal	PAT	Elongated	Lethal	PAT	Elongated	
<i>D1086.2</i>	0	0	48	7	6	39	1.074
ceh-49	0	2	91	8	5	12	1.162
<i>hlh-19</i>	0	0	12	5	0	12	0.950
<i>hmg-1.2</i>	7	0	110	13	1	42	0.967
<i>hnd-1</i>	0	2	76	1	1	16	0.980
<i>mex-3</i>	135	0	0	13	0	0	0.950
nhr-134	4	0	65	11	4	10	1.131
<i>nhr-60</i>	0	7	135	4	0	8	0.903
<i>nhr-63</i>	3	2	136	5	0	9	0.937
<i>oma-2</i>	12	3	116	2	1	19	0.972
<i>T03E6.3</i>	17	0	30	22	3	16	1.025
<i>taf-5</i>	0	1	145	9	6	32	1.082
tbp-1	0	0	32	12	19	61	1.197

ceh-51	0	0	132	3	3	10	1.169
---------------	----------	----------	------------	----------	----------	-----------	--------------

FIGURES

IA



IB

	Pharyngeal Muscle	Gonad	Body Wall Muscle	non-Striated Muscle		
	Tinman Family	HAND Family	MRF Family	MADS Family	TWIST Family	T-box Family
Mouse	Nkx2-5	HAND1 HAND2	MyoD, Myf5, MRF4, Myogenin	Srf Mef2	Twist1	Tbx1 Tbx6
Worm	ceh-22	hnd-1	hnh-1	unc-120	hnh-8	mls-1
Fly	Tinman		nautilus	Dmef2	twi	org-1

Figure 1: Proposed interactions of transcription factors and their orthologs

(A) The proposed interactions of different myogenic transcription factors, as has been studied primarily in the C-lineage. (B) The known myogenic transcription factors in *C. elegans* and their orthologs in both insects and vertebrates. The MRF family, MADS family, and Twist family of transcription factors seem to be functionally conserved while the HAND family appears to have a somewhat divergent function.

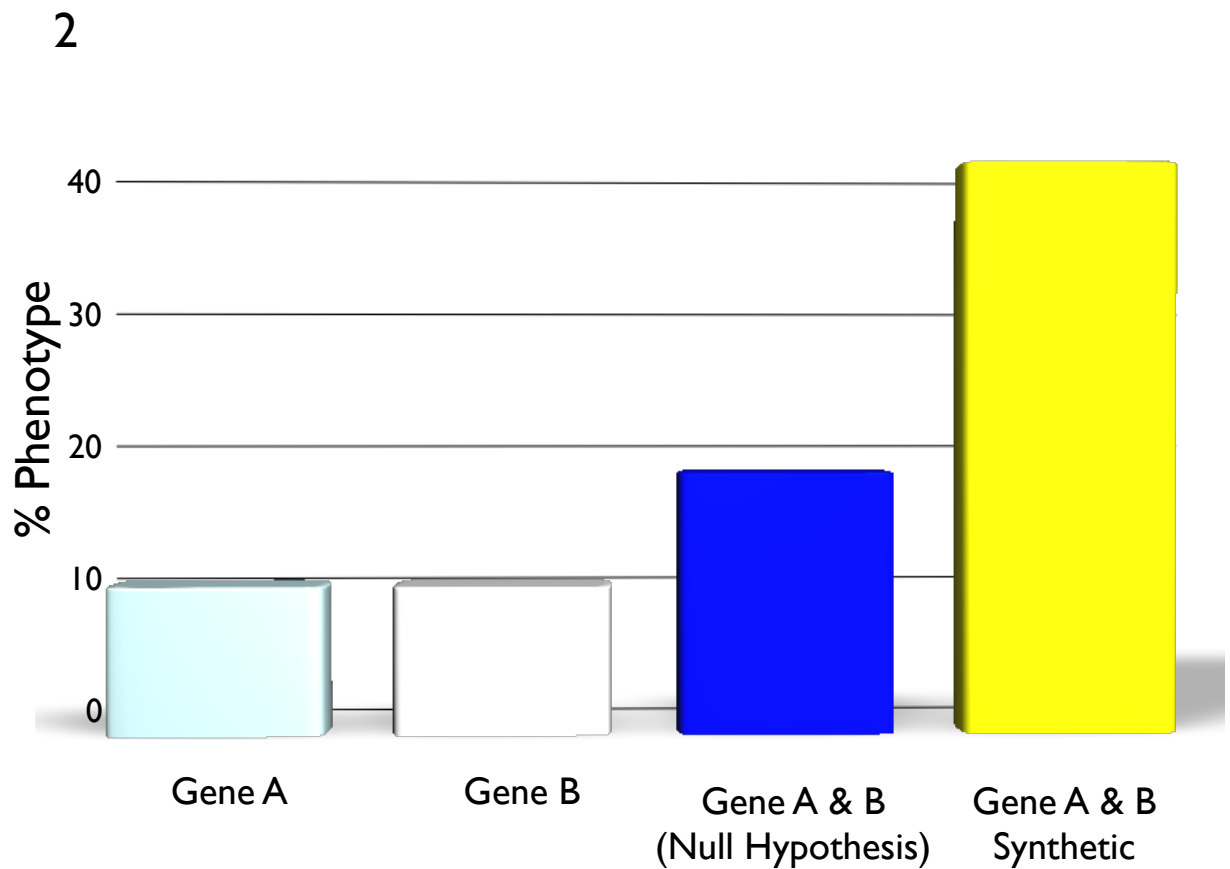


Figure 2: Synthetic Phenotypes

The compounded effect of transcription factor mutations in different overlapping pathways can lead to what is known as a synthetic interaction. The mutation of either gene may cause a small effect on lethality or other phenotypes (shown here in white). By compounding the two mutations, the lethality may be multiplicative, meaning that if 90% of the animals survived in each mutation alone, by chance independent mutations would lead to about 81% of the animals still surviving (shown in blue). However, if that survival drops precipitously, a significant increase in the mutant phenotype would indicate a

synthetic interaction (shown in yellow). Such increases indicate that though a pathway was buffered in the case of a single mutation, a second mutation causes the buffering pathway to collapse and a significant increase in the phenotype results.

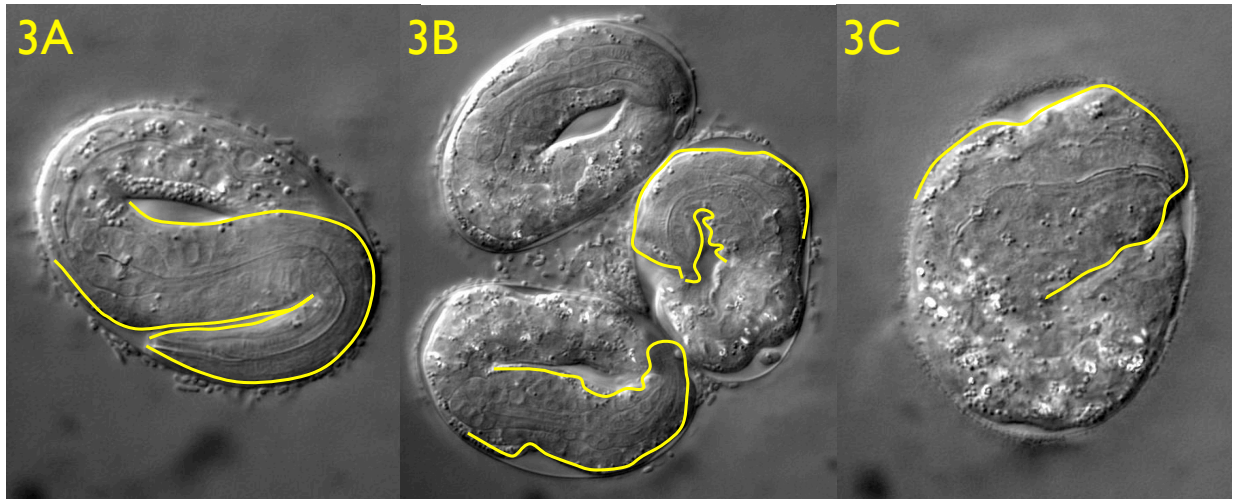


Figure 3: PAT and Lumpy Dumpy mutations

When muscle development is incomplete it has a profound impact on the morphology of the worm. Three phenotypes are shown with the head (determined by the length of the pharynx) highlighted in yellow for comparison. (A) Wild type animals completed elongation and have healthy muscle morphology, leading to a long, skinny worm. (B) If muscle differentiates defectively, lumpy dumpy animals arise. The muscle is not strong enough to properly elongate the worm and is generally uneven. These animals are capable of movement, but it is relatively uncoordinated. They are shorter and stouter than wild type animals, as illustrated by the size of the head. The mutation is generally lethal in the early larval stages. (C) Paralysis at the two-fold stage (PAT) animals do not have differentiated muscle. They are not able to elongate past this stage and retain a horseshoe shape. They are severely dumpy animals, as illustrated by the compactness of the head. These animals are completely incapable of movement except for their pharynx. The

mutation is lethal either in the embryo or in the early larval stage, as movement is not necessary for hatching.

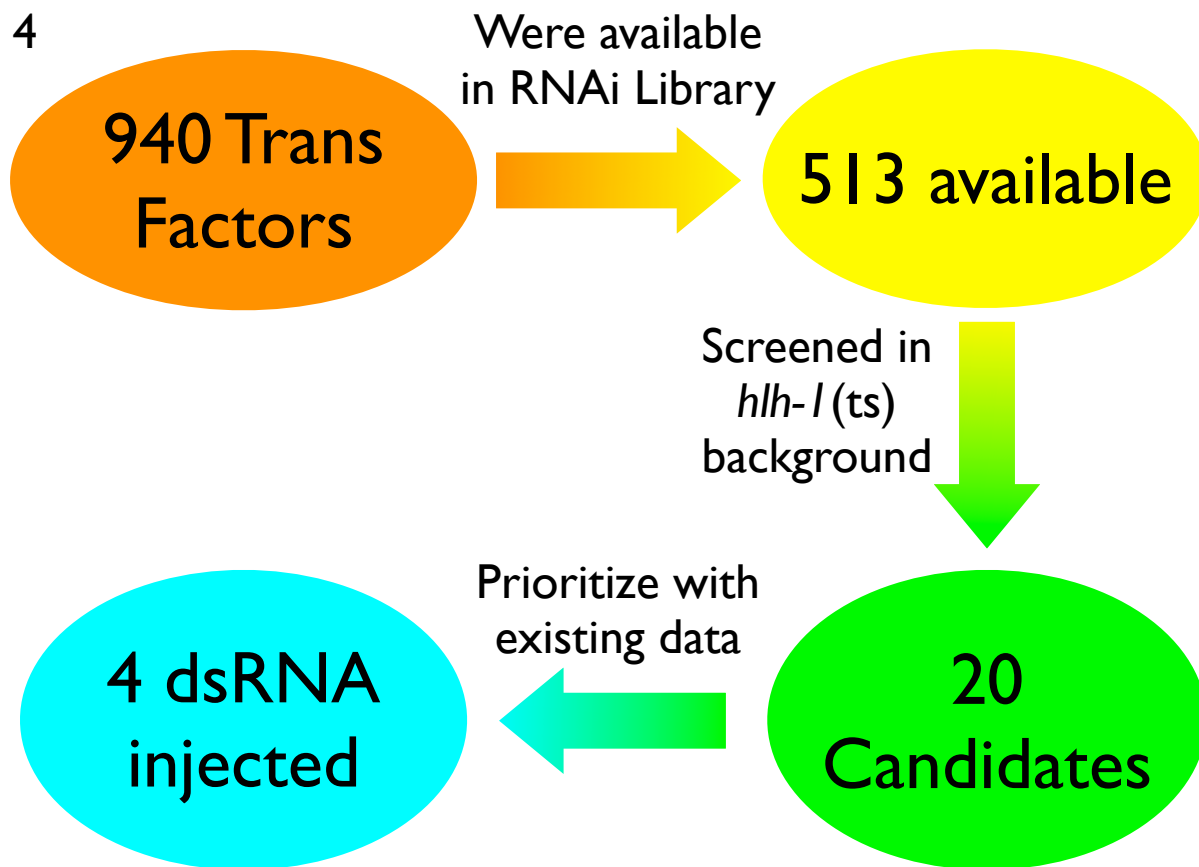


Figure 4: Experimental Plan

Initially all transcription factors available in the Ahringer RNAi Feeding Library were utilized for the initial feeding assay. Estimates on the total number of transcription factors ranges from 700 to 940 total factors. Based on initial results, 20 factors were selected for further study based on their synthetic lethality in the initial screen. Of these 20, 4 were selected for dsRNA injection to examine their interactions with other transcription factors in the myogenic network.

5

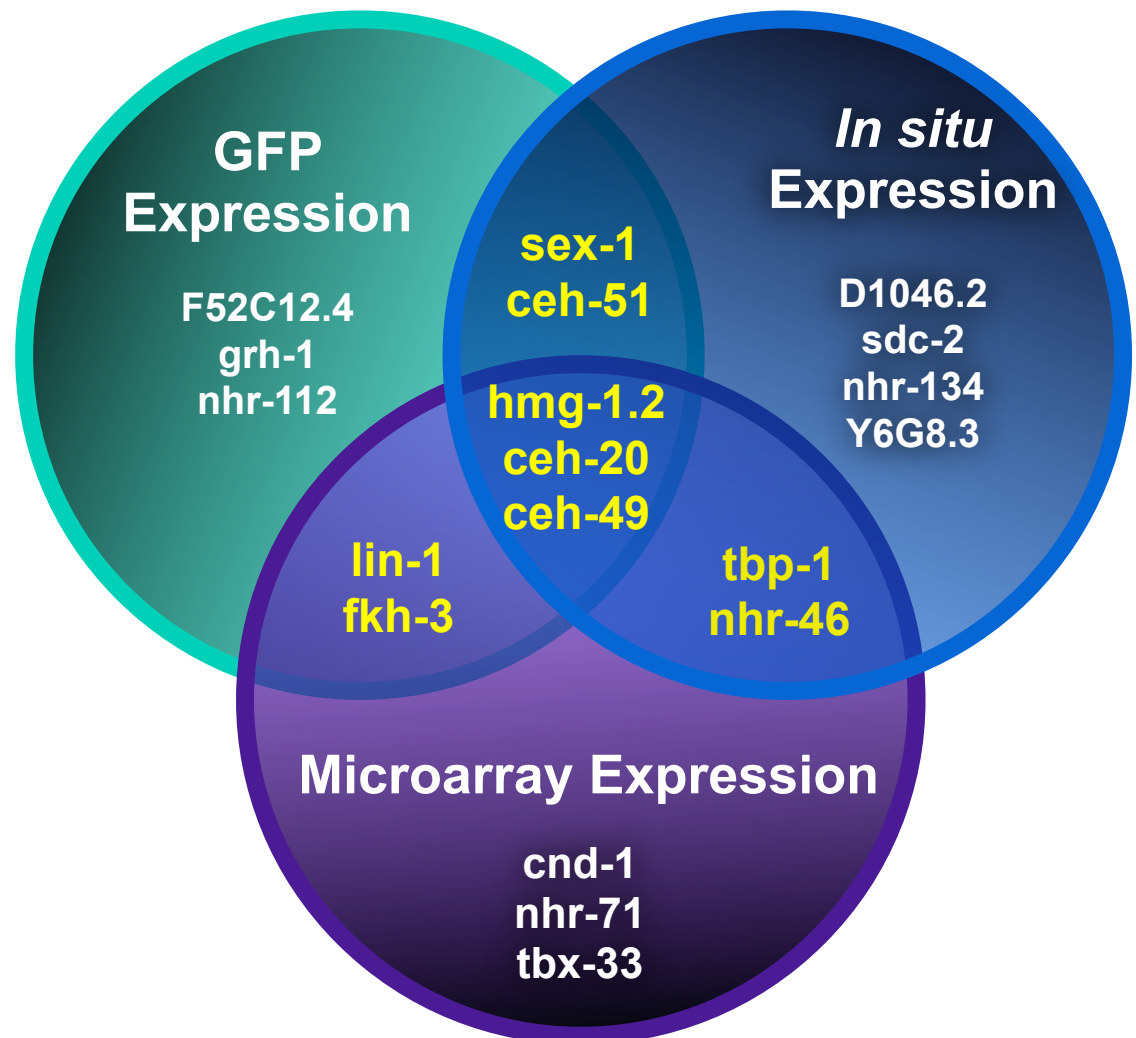


Figure 5: Categorization and ranking of top hits

The top hits from the synthetic lethal screen were ranked based on existing bioinformatic knowledge. Data was taken on where and when GFP reporters, microarrays, and *in situ* hybridization showed expression. GFP and *in situ* expression were considered to be more reliable than the microarray data due to the scale and precision of such experiments. Three genes were most favourably ranked due to three sets of data corroborating their

presence in embryonic muscle: *hmg-1.2*, *ceh-20*, and *ceh-49*. Two other genes were also considered as good candidates due to overlapping GFP expression and *in situ* data: *ceh-51* and *sex-1*.

6

RNAi

	none	<i>hlh-1</i>	<i>hnd-1</i>	<i>unc-120</i>	<i>ceh-51</i>
N2	0% n = 411	0% n = 107	0.9% n = 222	0.5% n = 198	0.6% n = 160
<i>hlh-1</i> (ts)	1.4% n = 634	12.5% n = 16	49.2% n = 65	30.8% n = 52	48.6% n = 70
<i>hnd-1</i>	1% n = 203	47.4% n = 38	0% n = 121	4.3% n = 69	16.2% n = 130
<i>unc-120</i> (ts)	0% n = 86	2% n = 51	44.4% n = 18	4.5% n = 22	30% n = 30
<i>ceh-51</i>	1% n = 194	47.2% n = 125	9.9% n = 213	10.4% n = 163	6.6% n = 166


No Interaction  Synthetic Interaction

Figure 6: dsRNA injection interactions across multiple backgrounds

To further study the interactions of the candidate transcription factors with the different myogenic factors, some of the genes were injected in different mutant backgrounds. Known interactions were replicated as controls. The mutant *hlh-1(cc561)* had strong interactions with all the injected RNAi constructs. Some interactions, such as those between *unc-120* and *hnd-1*, supported by (Baugh and Hunter, 2006), and *unc-120* and

ceh-51 were not symmetrical, suggesting that the strength of the mutations and RNAi were not equivalent.

7

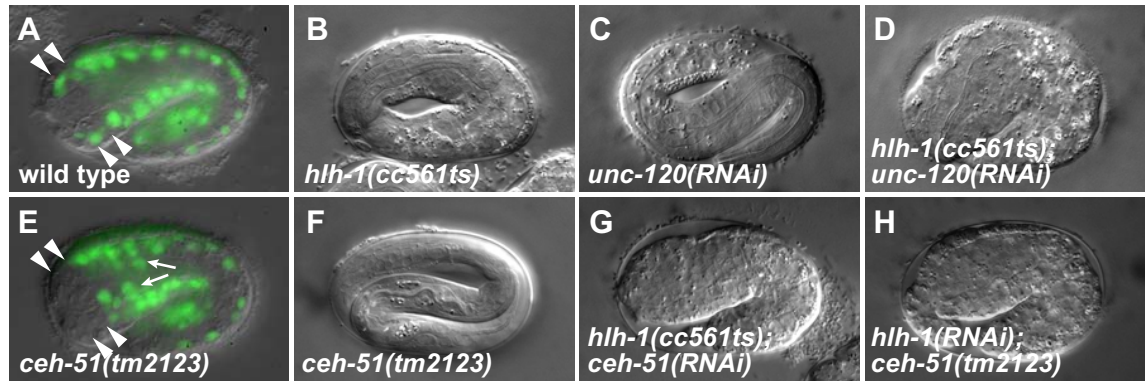


Figure 7: Synthetic PAT

Images of the animals after different interactions are shown. (A, E) Animals marked with *unc-120::GFP* are shown in wild type and in the *ceh-51(tm2123)* mutant. Additional expression is seen in the mutant. (B, C, F) Each mutation by itself or RNAi against wild type will not lead to a PAT phenotype. (D, G, H) However, the combination does produce a synthetic PAT phenotype.

8

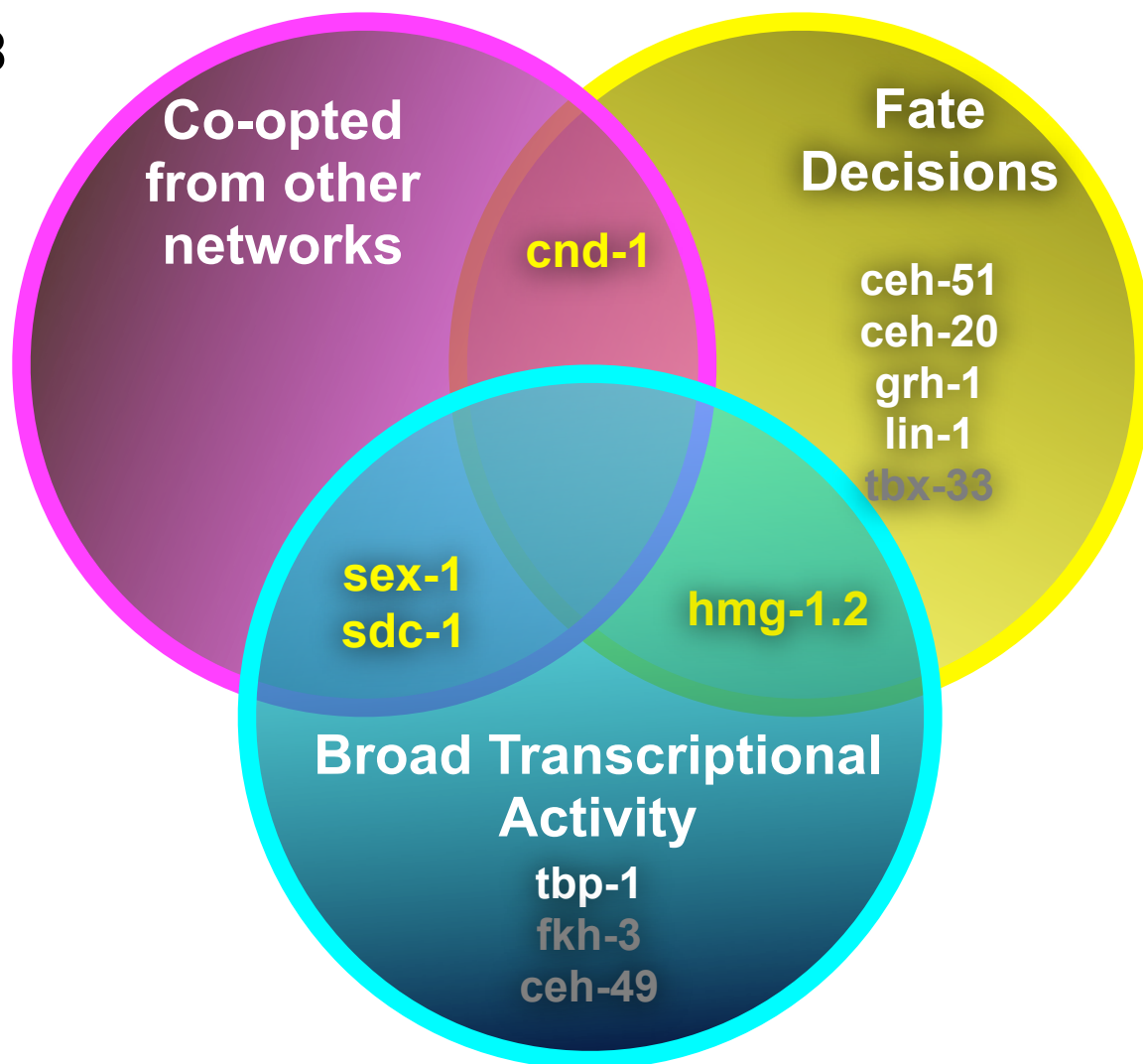


Figure 8: Categories of genes giving a synthetic PAT phenotype

The genes identified in this screen that have been previously observed and described fall loosely into three main categories: Genes involved in fate decisions within the muscle, broadly expressed genes involved in general transcription, and regulators of other transcriptional networks. It is possible that the transcription factors involved in fate decision within the muscle network are needed to help buffer terminal target gene expression in mutant muscle cells. Likewise, the presence factors, such as *ceh-20* and

grh-1, involved in Hox regulation and activation could explain the presence of genes involved in sex determination/dosage compensation (*sdc-1* and *sex-1*) and neuron development (*cnd-1*). These genes therefore may be some of the genes activated to buffer downstream target activation. General transcriptional frailty of the mutant explains the presence of genes broadly expressed and critical for proper transcriptional function, which includes both *tbp-1* and the sex determinatin/dosage compensation genes.

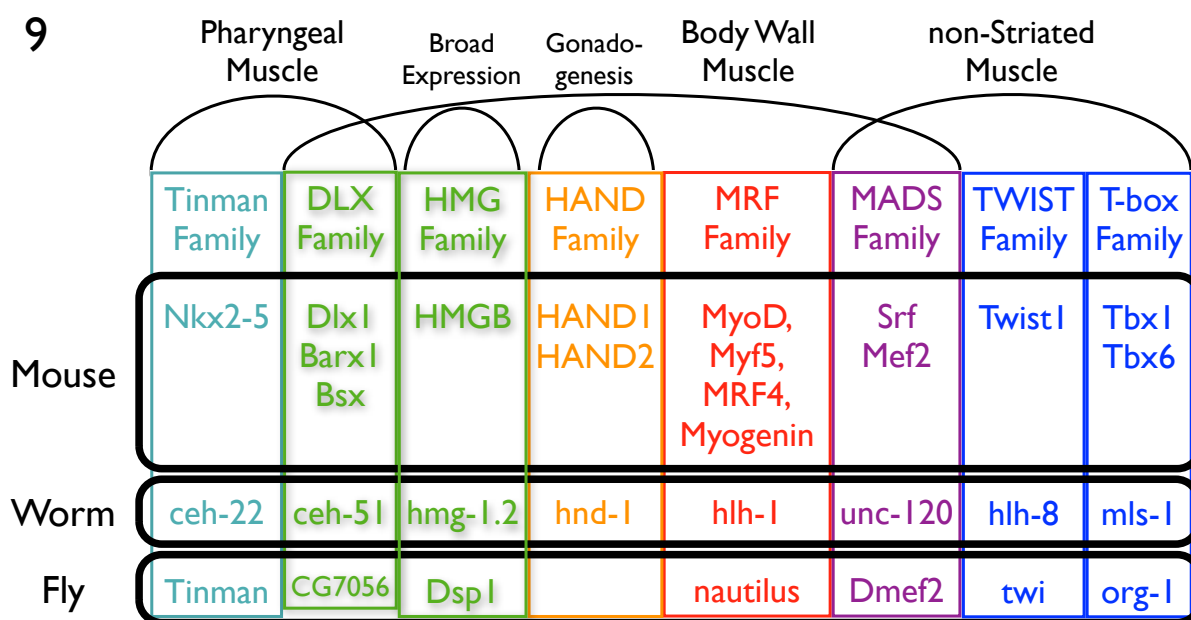


Figure 9: Expanded myogenic family

The expansion of groups of transcription factors involved in myogenesis is shown. The association of some categories with muscle may be unique to the nematodes, such as with *hnd-1*, due to the peculiarities of lineage specification. However, in some cases the orthologs may play important roles either in myogenesis in other phyla or play a parallel role in tissue specification. This study has added the HMG and DLX families of transcription factors to the set of myogenically important genes.

REFERENCES

Baugh, L.R., Hill, A.A., Claggett, J.M., Hill-Harfe, K., Wen, J.C., Slonim, D.K., Brown, E.L., and Hunter, C.P. (2005a). The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the *C. elegans* embryo. *Development* *132*, 1843-1854.

Baugh, L.R., and Hunter, C.P. (2006). MyoD, modularity, and myogenesis: conservation of regulators and redundancy in *C. elegans*. *Genes Dev* *20*, 3342-3346.

Baugh, L.R., Wen, J.C., Hill, A.A., Slonim, D.K., Brown, E.L., and Hunter, C.P. (2005b). Synthetic lethal analysis of *Caenorhabditis elegans* posterior embryonic patterning genes identifies conserved genetic interactions. *Genome Biol* *6*, R45.

Broitman-Maduro, G., Maduro, M.F., and Rothman, J.H. (2005). The noncanonical binding site of the MED-1 GATA factor defines differentially regulated target genes in the *C. elegans* mesendoderm. *Dev Cell* *8*, 427-433.

Broitman-Maduro, G., Owrighi, M., Hung, W.W., Kuntz, S., Sternberg, P.W., and Maduro, M.F. (2009). The NK-2 class homeodomain factor CEH-51 and the T-box factor TBX-35 have overlapping function in *C. elegans* mesoderm development. *Development* *136*, 2735-2746.

Fernandes, J.S., and Sternberg, P.W. (2007). The tailless ortholog *nhr-67* regulates patterning of gene expression and morphogenesis in the *C. elegans* vulva. *PLoS Genet* *3*, e69.

Fraser, A.G., Kamath, R.S., Zipperlen, P., Martinez-Campos, M., Sohrmann, M., and Ahringer, J. (2000). Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* *408*, 325-330.

Fukushige, T., Brodigan, T.M., Schriefer, L.A., Waterston, R.H., and Krause, M. (2006). Defining the transcriptional redundancy of early bodywall muscle development in *C. elegans*: evidence for a unified theory of animal muscle development. *Genes Dev* *20*, 3395-3406.

Grove, C.A., De Masi, F., Barrasa, M.I., Newburger, D.E., Alkema, M.J., Bulyk, M.L., and Walhout, A.J. (2009). A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* *138*, 314-327.

Harfe, B.D., Branda, C.S., Krause, M., Stern, M.J., and Fire, A. (1998a). MyoD and the specification of muscle and non-muscle fates during postembryonic development of the *C. elegans* mesoderm. *Development* *125*, 2479-2488.

Harfe, B.D., and Fire, A. (1998). Muscle and nerve-specific regulation of a novel NK-2 class homeodomain factor in *Caenorhabditis elegans*. *Development* *125*, 421-429.

Harfe, B.D., Vaz Gomes, A., Kenyon, C., Liu, J., Krause, M., and Fire, A. (1998b). Analysis of a *Caenorhabditis elegans* Twist homolog identifies conserved and divergent aspects of mesodermal patterning. *Genes Dev* *12*, 2623-2635.

Hill, A.A., Hunter, C.P., Tsung, B.T., Tucker-Kellogg, G., and Brown, E.L. (2000). Genomic analysis of gene expression in *C. elegans*. *Science* *290*, 809-812.

Hunter, C.P. (1999). Genetics: a touch of elegance with RNAi. *Curr Biol* *9*, R440-442.

Jiang, L.I., and Sternberg, P.W. (1999). An HMG1-like protein facilitates Wnt signaling in *Caenorhabditis elegans*. *Genes Dev* 13, 877-889.

Jiang, Y., Shi, H., and Liu, J. (2009). Two Hox cofactors, the Meis/Hth homolog UNC-62 and the Pbx/Exd homolog CEH-20, function together during *C. elegans* postembryonic mesodermal development. *Dev Biol* 334, 535-546.

Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N., and Davidson, G.S. (2001). A gene expression map for *Caenorhabditis elegans*. *Science* 293, 2087-2092.

Lehner, B., Crombie, C., Tischler, J., Fortunato, A., and Fraser, A.G. (2006). Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nat Genet* 38, 896-903.

Liu, H., Strauss, T.J., Potts, M.B., and Cameron, S. (2006). Direct regulation of *egl-1* and of programmed cell death by the Hox protein MAB-5 and by CEH-20, a *C. elegans* homolog of Pbx1. *Development* 133, 641-650.

Liu, X., Long, F., Peng, H., Aerni, S.J., Jiang, M., Sanchez-Blanco, A., Murray, J.I., Preston, E., Mericle, B., Batzoglou, S., *et al.* (2009). Analysis of cell fate from single-cell gene expression profiles in *C. elegans*. *Cell* 139, 623-633.

Mathies, L.D., Henderson, S.T., and Kimble, J. (2003). The *C. elegans* *Hand* gene controls embryogenesis and early gonadogenesis. *Development* 130, 2881-2892.

Nusbaum, C., and Meyer, B.J. (1989). The *Caenorhabditis elegans* gene *sdc-2* controls sex determination and dosage compensation in XX animals. *Genetics* 122, 579-593.

Potts, M.B., Wang, D.P., and Cameron, S. (2009). Trithorax, Hox, and TALE-class homeodomain proteins ensure cell survival through repression of the BH3-only gene *egl-1*. *Dev Biol* 329, 374-385.

Reece-Hoyes, J.S., Deplancke, B., Shingles, J., Grove, C.A., Hope, I.A., and Walhout, A.J. (2005). A compendium of *Caenorhabditis elegans* regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biol* 6, R110.

Reece-Hoyes, J.S., Shingles, J., Dupuy, D., Grove, C.A., Walhout, A.J., Vidal, M., and Hope, I.A. (2007). Insight into transcription factor gene duplication from *Caenorhabditis elegans* Promoterome-driven expression patterns. *BMC Genomics* 8, 27.

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

Tabara, H., Motohashi, T., and Kohara, Y. (1996). A multi-well version of in situ hybridization on whole mount embryos of *Caenorhabditis elegans*. *Nucleic Acids Res* 24, 2119-2124.

Venkatesan, K., McManus, H.R., Mello, C.C., Smith, T.F., and Hansen, U. (2003). Functional conservation between members of an ancient duplicated transcription factor family, LSF/Grainyhead. *Nucleic Acids Res* 31, 4304-4316.

Williams, B.D., and Waterston, R.H. (1994). Genes critical for muscle development and function in *Caenorhabditis elegans* identified through lethal mutations. *J Cell Biol* 124, 475-490.

Chapter 4: Genome-wide studies of HLH-1 regulation and binding in *C. elegans* myogenesis reveal regulatory interaction between the body wall and non-striated muscle gene networks

Genome-wide studies of HLH-1 regulation and binding in *C. elegans* myogenesis reveal regulatory interaction between the body wall and non-striated muscle gene networks.

Steven G. Kuntz^{1,2}, Brian Williams¹, Lorian Schaeffer¹, Paul W. Sternberg^{1,2*}, Barbara J. Wold^{1*}

1 Division of Biology and 2 Howard Hughes Medical Institute, California Institute of Technology, Pasadena, CA 91125, U.S.A.

* To whom correspondence should be addressed. Email: woldb@caltech.edu, pws@caltech.edu

Keywords: hlh-1, muscle differentiation, *C. elegans*, RNA-seq, ChIP-seq, cis-regulation

ABSTRACT

C. elegans body wall muscle differentiation is resilient to mutation of each major transcription factor in its core myogenic network, including HLH-1/MRF, the ortholog of vertebrate MyoD/myogenin and *Drosophila* Nautilus. This apparent robustness to loss of function raises questions about the specific role of HLH-1 and about the underlying network structure. We identified 2175 genes preferentially expressed in body wall muscle by using an RNAi knock down design to increase the proportion of worm specified as muscle. The impact of *hlh-1* mutation on global gene expression, quantified by RNA-seq, showed that 10% (216) of the muscle genes and 662 widely expressed genes depend significantly on HLH-1. HLH-1 binding was detected by chromatin immunoprecipitation sequencing (ChIP-seq) at 9447 sites in the genome, with 67% of HLH-1 dependent genes

having one or more binding sites. HLH-1 binding was also widespread near other genes, marking 8315 loci (32% of all genes not specific to muscle) and suggesting that at most sites HLH-1 occupancy alone has little regulatory impact on the adjacent promoter. HLH-1 occupancy was associated with several motifs, including two previously described E-boxes, a novel binding motif, and several accessory motifs. A small group of 307 genes was significantly up-regulated in the *hlh-1* mutant, including transcription factors *hlh-8/twist* and *mls-1/tbx1*, which are known regulators of non-striated (sex-specific and enteric) muscle differentiation. This supports a model in which the impact of *hlh-1* mutation is dampened by myogenic factors shared by both muscle types, such as UNC-120, and by up-regulated “compensatory factors” whose expression is normally restricted to the non-striated muscle differentiation network which shares target genes with body wall muscle.

INTRODUCTION

The adult nematode has multiple types of muscle, including pharyngeal muscle, enteric muscles, sex specific muscles, and the body wall muscles. Nematode muscle shares a common origin with vertebrate and insect muscle and an evolutionarily ancient regulatory system {Fukushige, 2006 #18}. It is unknown when the different types of muscle diverged, but certain parts of their regulation are held in common. The primary focus of this study is embryonic nematode body-wall muscle, or BWM. Being functionally analogous to the skeletal muscle of vertebrates and insects {Chen, 1994 #17;Fukushige, 2006 #18;Albertson, 1976 #57}, BWM is responsible for locomotion and

is the most prominent muscle tissue in the animal by cell number and mass (81 embryonic and 14 post-embryonic muscle cells) {Sulston, 1977 #56;Sulston, 1983 #24}. Non-straited muscles, or NSM, are a distinct muscle group that consists of the anal depressor cell, the anal sphincter, the enteric muscles, and the sex specific muscles. NSM comprise a relatively minor fraction of the worm (4 embryonic and 16 post-embryonic muscles) {Sulston, 1977 #56;Sulston, 1983 #24}. The developmental lineages for the embryonic muscles are independent from the post-embryonic lineage {Sulston, 1983 #24}, which depends on different factors {Corsi, 2000 #48;Dichoso, 2000 #58;Harfe, 1998 #11;Harfe, 1998 #10;Krause, 1992 #46}. Developmental regulation of the two muscle types depends on different bHLH transcription factors and on a shared regulator, UNC-120, from the MADS family (Figure 1A). The molecular level relationship of BWM and NSM networks is a second focus of this study. The third major muscle type, pharyngeal muscle, is a pulsating muscle, possibly analogous to vertebrate and insect heart muscle {Okkema, 1994 #64;Haun, 1998 #65}, which uses different core regulators and is not a topic of this work.

BW muscle depends strongly on a pair of transcription factors: HLH-1 (CeMyoD) and UNC-120 (SRF), with both CEH-51 and HND-1 playing important early supporting roles {Fukushige, 2006 #18;Yanai, 2008 #1;Broitman-Maduro, 2009 #63}. All four genes can convert early blastomeres to muscle, with HLH-1 being the most efficient {Fukushige, 2006 #18;Fukushige, 2005 #19} and the only one to be expressed exclusively in body wall muscle and its progenitors {Chen, 1992 #55}. In each case, over-expression of one factor induces, either directly or indirectly, expression of HLH-1 and UNC-120. Despite their individual sufficiency for initiating myogenesis, loss of

function mutations have shown that no single factor is necessary for myogenesis {Baugh, 2005 #3;Fukushige, 2006 #18;Broitman-Maduro, 2009 #63}. This apparent robustness to mutation of nematode myogenesis has been interpreted as partial ‘redundancy’ or ‘compensation,’ but these are properties whose molecular details are unknown and which this study aims to better define at the whole genome level.

Both HLH-1(CeMyoD) and UNC-120 are thought to be direct transcriptional regulators of a few well-studied body wall muscle differentiation genes in the worm, such as *myo-3*, *unc-54*, and *pat-3* {Fukushige, 2006 #18;Francis, 1985 #204}. This appears analogous to their vertebrate and insect orthologs, the bHLH MRFs (MyoD and paralogs in vertebrates, Nau in *Drosophila*) and SRF/MEF2A,C,D in vertebrates and dMEF2 in *Drosophila* (Figure 1A). *hlh-1* in *C. elegans*, like its orthologs, is a dedicated myogenic factor expressed solely in BWM and its progenitors {Baugh, 2003 #5;Baugh, 2006 #2;Chen, 1992 #55;Chen, 1994 #17;Fukushige, 2006 #18;Fukushige, 2005 #19;Williams, 1994 #20;Yanai, 2008 #1}. HLH-1 RNA expression is first detected at the 28-cell stage {McGhee JD, 1992 #45}, although the expression is not strong and stable until the 90-cell stage {Krause, 1992 #46}. *Unc-120* is a dedicated myogenic factor, but is expressed in both the BWM and NSM {Baugh, 2005 #4;Baugh, 2005 #3;Fukushige, 2006 #18;Lei, 2009 #34;Williams, 1994 #20;Yanai, 2008 #1}. UNC-120 RNA expression is seen in the early embryo with HLH-1 within the first 2 hours of development {Baugh, 2005 #3;Dichoso, 2000 #58;Fukushige, 2006 #18}.

At the gene circuit level, there is an apparent analogy between worm and vertebrate myogenic regulation with the bHLH myogenic factor HLH-1 positively autoregulating {Lei, 2009 #34} and cross regulating the MADS factor UNC-120 to form

a positive feed-forward circuit {Yanai, 2008 #1}. HND-1 and CEH-51 activate this transcriptional circuit early in differentiation and apparently play a more limited and indirect role in differentiation {Fukushige, 2006 #18;Yanai, 2008 #1;Broitman-Maduro, 2009 #63}.

The core gene network for non-striated muscle includes predominately the bHLH factor *hlh-8* (an ortholog of vertebrate and insect Twist), together with *unc-120*, which is shared with the BWM regulatory network {Corsi, 2000 #48;Harfe, 1998 #10;Hunt-Newbury, 2007 #62;Liu, 2000 #40} and *mIs-1* (orthologous to vertebrate TBX1), which is used in a subset of the NSM {Kostas, 2002 #67;Reece-Hoyes, 2007 #70}. Analogous to HLH-1 in the BWM, HLH-8 expression is dedicated to NSM and its progenitors {Corsi, 2000 #48;Harfe, 1998 #10}. Pertinent to this study, HLH-8 expression overlaps transiently with HLH-1 in the M-lineage cells whose progeny go on to produce 14 BWM cells expressing only HLH-1, 16 NSM cells expressing only HLH-8, and two non-muscle coelomocytes {Sulston, 1977 #56}. When ectopically expressed, HLH-8 can produce NSM phenotypes in other cell types that normally do not express HLH-1 {Harfe, 1998 #10;Zhao, 2007 #68;Wang, 2006 #69}.

Until this work, the locations of HLH-1 protein binding in vivo was known only for a few specific candidate sites in the worm genome, and these were predicted to bind HLH-1 because they are adjacent to BW-muscle specific genes {Lei, 2009 #34}. Recent genome-wide studies of binding by the mouse orthologs, MyoD1 {Cao, 2010 #85} and myogenin (Wold et al., in preparation), surprisingly found that the number of sites occupied in the mammalian genome for these functionally dedicated factors is unexpectedly high (15,000 - 80,000). A similar study in *C. elegans* using PHA-4

similarly uncovered thousands of sites throughout the genome {Zhong, 2010 #207}. These sites of occupancy are located near a majority of genes in the vertebrate genome, rather than being specifically adjacent to skeletal muscle-specific genes, as might have been naively expected for dedicated muscle factors. These findings from a large genome raise a series of questions about what characteristics of bound regions determine the regulatory action - or lack thereof - by myogenic bHLH factors. The worm, with its smaller genome, presents the opportunity to learn whether myogenic factor occupancy is correspondingly numerous and widespread in the more compact worm genome. Worm genetics further affords direct identification of HLH-1 regulatory targets and evaluation of the relationship between HLH-1 in vivo occupancy and observable regulatory dependency. Such analyses are made difficult in vertebrates by the presence of four paralogous MRFS and four muscle MADS factors with partially overlapping functions.

A technical challenge for functional genomic studies of worm myogenesis is that BWM comprises only ~12% of the embryo {Sulston, 1983 #18666}, which means that genome-wide biochemical assays such as ChIP and transcriptome quantification are complicated by contamination from the remaining 88% of cells. Nevertheless, nematodes offer great advantages in genetic manipulation and understanding of muscle development, some of which can be used to experimentally ameliorate the cellular impurity problem. Specifically, RNAi feeding knockdown, in which bacteria expressing a double-stranded gene specific RNA are fed to the worms to knock down a target gene, can be used to suppress genes critical for lineage selection. This causes more cells to adopt a muscle fate {Baugh, 2005 #4}. Due to the deterministic cell lineage of *C. elegans*, knocking down individual genes can significantly change the cellular make-up

of the animal. By knocking down *mex-3* in the embryo, PAL-1 continues to be expressed in the AB lineage, causing it to divide twice, and each granddaughter to divide like the C lineage. From this change, rather than producing 1 body wall muscle cell, 3 enteric muscle cells, and most of the pharynx in the normal AB lineage, it instead produces ~80 body wall muscle cells {Sulston, 1983 #24, Draper, 1996 #22, Hunter, 1996 #28}. Similarly, a knockdown of *skn-1* will prevent the EMS lineage from producing its normal range of fates and instead it too will adopt a C-like fate {Bowerman, 1992 #23, Blackwell, 1994 #29}, thus preventing formation of the final enteric muscle and the M-cell lineage {Sulston, 1983 #24}. Finally, knocking down *elt-1* – the master regulator for hypodermal specification – will cause the remaining hypodermal cells in the C-lineage to adopt a mesodermal muscle specification {Michaux, 2001 #27}. The possibility of varying the degree of conversion to C-lineage muscle, by performing single or triple RNAi, is used here to help identify and interpret differences in signal strength and quality compared with each other and with the N2 wild type.

In this study (Figure 1B), we use wild type N2 and single- and triple-RNAi muscle-enriched worms to identify genes with strong muscle preferential expression versus genes expressed more widely in the animal. We then determine by RNA-seq transcriptome analysis which genes from both groups are targets of HLH-1 regulation, including both direct and indirect targets. A majority of BW muscle-specific genes are down-regulated by *hlh-1* mutation. Among genes up-regulated in HLH-1 mutants, we identify a set of transcription factors known to positively regulate NS muscle differentiation and discuss implications of this finding for explaining the tolerance of worm myogenesis to *hlh-1* mutation. Finally, we determine HLH-1 protein occupancy

across the genome by ChIP-Seq, and evaluate how physical targets of HLH-1 are related to both regulatory targets and DNA binding motifs for HLH-1, HLH-8 and candidate accessory factor motifs.

RESULTS

RNA-seq

Only one-sixth of worm cells normally become body wall muscle and this low fraction presents a signal to noise problem for transcriptome and whole genome assays. We addressed this by using previously established genetic manipulations to increase the fraction of cells specified to become body wall muscle. There are two different strategies known to increase the proportion of muscle. We specifically avoided an *hlh-1* overexpression design, because that would alter the *hlh-1* muscle differentiation regulatory circuit itself {Fox, 2008 #36} and would likely skew *hlh-1* expression to non-physiological concentrations, leading to uncertain changes in the composition, expression, and behavior of target genes. Instead, we generated worms with more muscle by manipulating the cell lineage specification prior to the onset of HLH-1 expression and activity. We accomplished this by RNAi knock down of specification genes, as suggested by previous studies: *mex-3* plays a role 3 cell divisions before HLH-1 expression {Draper, 1996 #22;Hunter, 1996 #28}, *skn-1* plays a role 2 or 3 cell divisions beforehand {Blackwell, 1994 #29;Bowerman, 1992 #23}, and *elt-1* plays a role around the time *hlh-1* would be activated {Michaux, 2001 #27;Spieth, 1991 #49}, but still permits *hlh-1* expression. We utilized three conditions for all of our analyses: no RNAi (the bacteria

contain an empty vector), *mex-3* RNAi, and *elt-1*, *mex-3*, and *skn-1* triple RNAi. Because the triple RNAi produces almost entirely muscle it is likely to have the highest signal to noise ratio {Baugh, 2005 #4;Bowerman, 1992 #23;Draper, 1996 #22;Page, 1997 #21}. However, knocking down multiple genes via RNAi can significantly reduce the penetrance and overall effect {Gonczy, 2000 #42}. As this reduction in efficiency has been postulated to be an uptake issue, we concatenated the RNAi transcripts of the three genes to assure unified action rather than a stochastic mixed population. The advantage of having a high conversion to muscle in the triple RNAi sample is tempered by the fact that it is dominated by the C-lineage. Therefore, we also included the *mex-3* RNAi worms, as they have a significant amount of muscle from the EMS and D-lineages, in addition to having twice the body wall muscle of wild type animals {Draper, 1996 #22}. This provides a sampling of body wall muscle lineages, as well as a graded series in concentration of muscle nuclei.

We performed RNA-seq in wild type N2 and the temperature sensitive *hlh-1(cc561)* mutant background to learn the regulatory target of HLH-1 and begin to understand how the system can compensate for its absence. The RNAi strategy succeeded in enriching for classical markers of BWM (Table 1; Supplementary Material).

To identify regulatory targets of HLH-1, both direct and indirect, we performed RNA-seq transcriptome profiling of polyA⁺ RNA {Mortazavi, 2008 #30} from *hlh-1(cc561)* mutant and N2 wild-type embryos. The developmental time point was selected to guarantee that cells had already been specified, thus capturing embryos in the process of differentiation to observe the expression of genes important in muscle development. Differences in RNA levels between mutant and wild-type were quantified in untreated,

mex-3 RNAi, and triple *mex-3/skn-1/elt-1* RNAi. This experiment is expected to identify both direct and indirect regulatory targets of *hlh-1* and to allow us to parse targets that are strongly muscle-enriched versus those widely expressed in both muscle and non-muscle cell types (Figure 1B).

Because HLH-1 has been shown to function as a direct activator at several muscle specific loci {Lei, 2009 #34}, a simple expectation is that additional direct targets across the genome will be dominated by down-regulation of the corresponding RNAs. We observed 878 candidate genes for positive regulation, based on significant reduction of RNA in *hlh-1(cc561)* embryos.

We expect some fraction of genes regulated by HLH-1 to be expressed exclusively in body wall muscle, just like HLH-1. However, it is possible that other genes regulated by HLH-1 within muscle will be expressed – under the regulatory control of different factors – in other cell types. To determine the distribution between muscle-specific and broadly expressed targets we separated out genes enriched in muscle-rich animals. Genes were identified where the expression level was higher in animals with RNAi-based muscle enrichment. Because the wild-type animals contain roughly 12% muscle, muscle-specific genes should still be present in the muscle-enriched animals. Therefore, we are looking for increases in expression, rather than presence/absence. Alternatively, when attempting to identify genes absent in muscle tissue, the corresponding decrease in expression between the muscle-enriched and wild type animals is much more severe. This is easily observed by genes such as that encoding the non-muscle troponin C, *tnc-2*, with expression several times higher in the muscle-normal animals (Figure 2C).

To identify an initial set of muscle-specific genes, the expression levels from the RNA-seq analysis was taken from wild-type animals through two approaches for muscle-enrichment and from two biological replicates of muscle-normal animals. Most important muscle genes are well within these bounds. From a set of muscle structural genes described by Fox et al. {Fox, 2008 #18752}, we identified 20 of the 38 genes in our sample. The remaining 18 were expressed but were not enriched in muscle due to their expression in the pharynx and other tissues. In total, 2,175 genes appear to be enriched in muscle (Figure 2A). Examples of genes that are preferentially expressed in muscles include *unc-54*, *myo-3*, *tnt-3*, *dhp-2*, etc. (Table 1). The main drawback of this technique is that there is no clear-cut way to identify genes that are expressed exclusively in muscle apart from those simply expressed primarily or preferentially in muscle. Conspicuously absent from this list includes genes such as the transcription factors *ceh-34*, *mef-2*, *unc-120*, and *hnd-1*. *unc-120* and *hnd-1* are known to be expressed elsewhere {Mathies, 2003 #61; Hunt-Newbury, 2007 #62}, while *ceh-34* and *mef-2* have different functions from their muscle-controlling homologues {Dichoso, 2000 #58; Dozier, 2001 #59; Amin, 2009 #60}.

Numerous genes are highly enriched outside of the muscle, as expected. From our observations, 3901 genes are expressed preferentially in non-muscle tissues (Figure 3; Supplementary Material). As expected, there are more genes expressed outside of muscle than within muscle. The actual number of genes is almost certainly higher, but because numerous genes are expressed in only a single cell or at lower levels, they are below the assay's threshold.

There exist several published sets of data that describe embryonic genes, both muscle and non-muscle {Fox, 2008 #36, Fukushige, 2005 #19, Fukushige, 2006 #18, Von Stetina, 2007 #18763}. Since the different datasets come from different experimental conditions and are acquired with different techniques, it is expected that there will not be perfect overlap between the sets, but the overlap should be statistically significant (Supplementary Material). We see a significant overlap of muscle-enriched genes with the existing muscle-enriched datasets from Fox et al. {Fox, 2008 #36} and Fukushige et al. {Fukushige, 2006 #18}. The set of genes enriched in non-muscle is statistically underrepresented in each of these lists, as expected. Of non-muscle datasets, such as Von Stetina et al {Von Stetina, 2007 #37} and Fox et al B {Fox, 2005 #39}, we see a statistically significant overlap with the non-muscle genes and little overlap with the muscle genes, again as expected. This supports our declaration of these genes as muscle genes.

We are interested in what genes depend on *hlh-1* expression. Therefore, by observing *hlh-1* temperature sensitive mutants we can identify which genes are most and least affected (Figure 3). One caveat of this technique is the comparison of muscle between the two samples. If overall expression of muscle cell genes is reduced in the mutant, non-muscle genes will have proportionally higher expression. Such changes will still be identified as non-muscle due to the muscle-enrichment comparisons. It appears that many muscle genes are not affected by the mutation, which bolsters our hypothesis that we can target certain genes as being directly or strongly but indirectly regulated by *hlh-1*. Such unaffected genes, genes that are below our threshold for a statistically significant decrease in expression in the *hlh-1* mutant muscle enriched samples, include

the major muscle myosins, actins, and many other genes involved in terminal muscle differentiation (Figure 3). These genes are good candidates for understanding what factors work with *hlh-1* to drive muscle differentiation. If other cis-regulatory elements are identified, then there are partially overlapping cis-regulatory elements while if only one cis-regulatory element is identified, then cooperative binding may lead to sustained functionality. It is possible that the expression of at least some of these genes will only be affected by a set of synthetic lethal mutations or knock-downs.

It appears that *unc-120* does not appear to be significantly negatively affected by the *hlh-1* mutation, in contrast to what was observed by Yanai, et al. {Yanai, #1}. This may be due to the nature of the mutation, rather than RNAi, or relate to observations being in the late embryo after *hnd-1* has been shut off rather than the early embryo when specification is still taking place.

Genes that are negatively affected by *hlh-1*'s absence include a number of expected genes, including *tnt-3*, a muscle troponin, and *srp-1*, a serine protease inhibitor. Other troponins, such as *tnt-1* are not affected, indicating redundant gene targets in the muscle. An impact is expected; however, none of these genes lose all expression. *dnp-2* experiences among the most severe losses in expression, but is still has baseline expression. Though significantly diminished, they are still expressed, indicating that other transcription factors still drive their expression. More interestingly, a number of genes that have not been previously described as relevant to muscle have reduced expression in the mutant muscle. Additionally, several predicted transcription factors are found in this group as well, including certain ribosomal proteins such as *rpl-2*, *rpl-4*, and *rps-6*.

Though a much smaller group, some genes are relatively higher in the mutant muscle. These 307 genes are certainly candidates for compensation following the loss of *hlh-1* and are muscle-enriched only in the mutant. In the wild-type animals, their expression is actually lower in muscle than the rest of the animal, though this is reversed in the mutant. Some of the most striking examples of upregulated genes are the transcription factors *hlh-8* and *mIs-1*. Though these are myogenic transcription factors, they function in the NSM, meaning their enrichment in embryonic bodywall muscle-enriched animals is surprising. The RNAi ensures that no expression from the enteric muscles or M cell is observed, as demonstrated in the wild type animals. Of the 307 genes upregulated in the mutant muscle, 96 have described expression patterns. Of those, 26 (27%) are expressed in *hlh-8* derived tissues: either the enteric muscle or sex-specific muscle (Table 2). We then took all genes described as being expressed in either BWM or NSM and looked at their overlap with expression levels. 63% of BWM genes are also expressed in NSM and 68% of NSM genes are expressed in BWM. However, only the group of genes expressed in BWM have a statistically significant overlap with *hlh-1* dependent gene expression. And only the group of genes expressed in NSM have a statistically significant overlap with genes enriched in the *hlh-1* mutant muscle. Therefore these genes may be part of a fate-switching compensatory apparatus. Though unexpected, this has been observed within the post-embryonic M-lineage {Harfe, 1998 #11} and analogous cross-network inhibition has been observed between muscle and epidermal networks {Yanai, 2008 #1}.

To address how this compensation may work, we looked at what activates the non-body wall muscle network, specifically *hlh-8* and *mIs-1*. *Ceh-20/Exd* is responsible

for the activation of *hh-8* {Jiang, 2009 #25;Liu, 2000 #40}, but usually works in conjunction with *unc-62/Meis1* {Jiang, 2009 #25;Potts, 2009 #26}. In the VC neuron, these genes can either repress *egl-1* if working with *lin-39* or activate *egl-1* if working with *mab-5* {Liu, 2000 #40, Potts, 2009 #26;Liu, 2000 #40;Potts, 2009 #26}. *Mab-5* is itself activated by *grh-1* in *Drosophila*, which is predicted to interact with *egl-15* in worm {Zhong, 2006 #33}. To further investigate *egl-15* expression, we looked at the different splicing variants, as EGL-15a is preferentially expressed in the sex myoblasts and vulval muscle and EGL-15b is expressed in body wall muscle {Kuroyanagi, 2007 #104}. RNA-seq data can be used to investigate splicing. Though the splices leading to exon 5B are unchanged in the mutant, there is a significant increase in the number of splices to exon 5A. There is a similar increase in splicing to the final 5 exons, which like exon 5A are specific to EGL-15a. This indicates that in the mutants, EGL-15a is significantly upregulated. Therefore the upregulated GRH-1 may interact with the vulval muscle version of EGL-15 to regulate MAB-5. In the post-embryonic SM cells where *hh-8* is normally active, proper cell migration and division is dependent on *mab-5* {Kenyon, 1986 #43}. This is of interest because *ceh-20*, *unc-62*, *lin-39*, *mab-5*, *grh-1*, and *egl-15* are all expressed in the *hh-1* mutant.

Synthetic PAT Screen

To further investigate these interactions, we performed a feeding RNAi synthetic paralysis-at-twofold (PAT) phenotype analysis of these transcription factors against the *hh-1(cc561)* mutant background. In nematodes, elongation in the embryo is dependent on contractile muscle to essentially squeeze the worm out. The PAT phenotype indicates that muscle differentiation has been halted. Though *unc-62* is always lethal, both *ceh-20*

and *grh-1* showed strong synthetic PAT phenotype with the *hlh-1* mutation (Table 3). This indicates that both of these genes may be critical for the continued differentiation of muscle in an *hlh-1* mutant, likely through the activation of the *hlh-8* and *mls-1* alternative differentiation pathway. Additionally, *nhr-63* showed a strong synthetic lethal phenotype. Though expressed in the NSM, its role is not known.

Since *unc-120* is necessary for the alternative muscle differentiation pathway, we wanted to see if this pathway is activated in an *unc-120* mutant. By looking at the mRNA expression levels of several genes, including *hlh-8* and *mls-1*, that are not elevated in the *unc-120* mutant, it is clear that the pathway is not similarly activated in these mutants and appears to be specific to the *hlh-1* mutation, as predicted.

Anti-HLH-1 ChIP-seq

In order to understand transcription factor behavior *in vivo*, we looked directly at transcription factor-DNA binding using ChIP-seq {Zhong, 2010 #207; Johnson, 2007 #212}. This may tell us how our predictions based on expression levels compare to transcription factor behavior. The *hlh-1(cc561)* mutation does not completely eliminate activity of the *hlh-1* gene and thus the *hlh-1* mRNA is still produced, though at a lower level than in wild-type worms (Figure 4A). This decreased level is due to non-sense mediated decay {Harfe, 1998 #11} and likely because of the collapse of the auto-regulatory loop at the permissive temperature. Because the parent generation was raised to the permissive temperature at the L4 stage prior to egg fertilization and embryogenesis, there is no residual functional HLH-1 in the embryo left over from the lower temperature. There is no maternal or zygotic requirement or effect {Chen, 1994 #17}. Using an

existing anti-HLH-1 antibody {Lei, 2009 #34}, we tried immunoprecipitation in the mutant. However, the antibody did not pick up any signal above background, even in the muscle-enriched animals. The mutation may destroy the epitope or the mutation is strong enough to prevent binding to chromatin. While the mutation is not a null and some transcription factor may still be present, it effectively destroys HLH-1 function {Lei, 2009 #34}, as seen in the collapse of autoregulation. To analyze the binding, peak intensities and locations were observed against background utilizing peak shifting to screen out noise {Pepke, 2009 #31} in each of the RNAi feeding conditions. However, the signal was extremely low in the muscle-normal animals, barely distinguishable from the background signal. The polyclonal antibody, though affinity purified, is possibly not strong enough or not selective enough to extract sufficient material for our assays in muscle normal wildtype animals. We obtained a set of targets from both the *mex-3* RNAi-fed animals (7032 targets) and the triple RNAi-fed animals (3452 targets). The intersection of these two sets was 1047 hits, which is greater than what would be the expected random overlap. Differences in binding between the two samples may derive from the different muscle content of the two conditions or from variability in binding at lower-level targets.

The HLH-1 bound target sequences are distributed across genes, and occur intergenically, upstream of the gene, within the exons, and within introns. Of the sites, 89% were found in intergenic DNA upstream of a known gene, 32% were within 500 bp 5'ward of the 5' start sequence, 16% in introns, 6.6% in coding exons, and 1.2% in UTRs (Table 4). Compared to random coverage of genomic regions, the intergenic, 500 bp proximal sequences, and the 5' UTR were enriched while the other regions are depleted

of binding sites. The signals in and near the proximal promoter region can be from two sources: direct binding of HLH-1 to promoter DNA captured by protein-DNA crosslink or indirect binding of HLH-1 captured by physical binding of HLH-1 – itself bound to a remote enhancer – to a promoter complex bound to the DNA. Protein:protein:DNA complexes are known to be retrieved under these conditions, which are also used for chromatin conformations and distant interaction studies {Fullwood, 2009 #102; Fullwood, 2009 #103}.

The genes that are nearest neighbors to the HLH-1 targets include numerous muscle-related genes as well as a number of undescribed genes. Associated genes include known muscle genes *unc-54*, *tnt-3*, *hlh-1*, *dhp-2*, etc. (Figure 5A). Both genes enriched in muscle and genes that are dependent on *hlh-1* are more likely to have HLH-1 binding either in the gene body or in the 5 kb upstream region than other genes (Table 5). Genes that are conspicuously absent include a number of genes where the binding site is at least a gene away downstream, as with *skr-2* and *skr-1* (Figure 5C) and *unc-120*, which has no observable binding. Other genes, such as *srp-1*, may have HLH-1 binding that simply falls below the observable threshold largely due to the high level of background (Supplementary Figure). Though *unc-62*, *mab-5*, and *grh-1* have HLH-1 binding, the gene *hlh-8* does not. In fact, if we look at all genes upregulated specifically in the *hlh-1* mutant muscle, there is a decrease in likelihood of them having an HLH-1 binding site nearby (Table 5).

To determine which motifs are enriched near the binding site we utilized multiple motif-finding algorithms on sequences within various radii of the binding site. Both a greedy motif-finding algorithm and MEME found similar motifs to be overrepresented in

the sequences, depending on the size of the radius utilized. This is not unexpected, given the statistical impact that varying the sequence volume has on motif-finders. The different motifs represent both the actual binding motif for the HLH-1 transcription factor and possible associated binding sites important for accessory transcription factor binding. By using a radius of 50 bp, the primary motif has been identified as an E-box (a motif with a CANNTG motif commonly bound by helix-loop-helix transcription factors) with the consensus sequence CAGCTG (Figure 4B). This matches with Grove et al.'s {Grove, 2009 #14} determination via in vitro yeast one-hybrid assays. Several additional motifs were identified that are indicative of GA or CT repeat-rich regions being important around the HLH-1 binding sites {Guhathakurta, 2002 #32, GuhaThakurta, 2004 #35}. These additional regions may relate to either degenerate binding sites for other associated transcription factors or perhaps markers for acetylation control of the surrounding chromatin. Additional motifs found using a 50 bp radius include a TCTGCG motif, the importance of which is unknown (Figure 4C).

With a radius of 100 bp, an additional E-box is identified: CAACTG (Figure 4D). This motif is predicted to be a secondary binding motif for HLH-1, identified previously both in vitro via yeast one-hybrid {Grove, 2009 #14} and by ChIP {Lei, 2009 #34}. The relative importance of each of these motifs is not known, but the CAGCTG motif is more prevalent among the identified transcription factor binding sites. As the radius is increased once more to 250 bp, the motif-finding algorithms no longer find the E-boxes, but do find two motifs that appear to be similar to previously identified muscle-related motifs: GAGACGCA (Figure 4E) and TCTCGCAA (Figure 4F) {Guhathakurta, #32}.

Although the motifs are identified with different sized input sequence, the location of the motifs in relation to the binding sites might be informative. By graphing the location of the binding sites we could identify the position-dependent nature of the motifs. The two E-box motifs and the TCTGCG motif are very position-dependent, being generally centered on the *hlh-1* binding site (Figure 4G), as expected for a transcription factor binding-motif. The other motifs are not nearly so position-dependent and are more evenly spaced throughout the observed ranges (Figure 4H).

Because *hlh-8* is upregulated in the mutant muscle, we decided to compare the prevalence of the HLH-8 binding site (CATATG) to the HLH-1 binding sites (CAGCTG) {Grove, 2009 #14}. As there is a very significant overlap between the genes known to be expressed in BWM and NSM, we wanted to see if HLH-8 E-boxes might co-localize with HLH-1 E-boxes. For both genes that are enriched in muscle and genes that are dependent on HLH-1 binding dependent, there are over twice as many HLH-1-specific E-boxes as HLH-8-specific E-boxes within a 250 bp radius of HLH-1 binding. This is consistent with the ratio of sites for non-HLH-1 dependent genes. Therefore the cis-regulatory elements are not significantly shared between factors, though genes may have multiple cis-regulatory elements to respond to each factor.

The prevalence of the motifs outside of the HLH-1 targeted regions varies between the different motifs. Several of the motifs, such as the E-boxes, are targeted by multiple transcription factors with non-overlapping expression patterns {Grove, 2009 #14, Krause, 1997 #13}. Given that these motifs depend on as little as 6 bases, it is possible that across 100 million random base pairs (the length of the *C. elegans* genome) that they could appear over 24,000 times by chance. Starting with a PSFM (position-

specific frequency matrix) as the reference motif, a 95% match guarantees an essentially perfect match (100% matches are generally impossible given the variation within the reference motif). An 85% match, depending on the transcription factor, may or may not be a real motif. A scan through the genome for matches to each of these motifs at 85% and 95% identity provides the baseline frequency (supplementary figure). The number of motifs identified within the anti-hlh-1 ChIP identified regions was also determined at the different thresholds. By comparing the different thresholds, interesting patterns emerge (Table 6). Though the total number of motifs decreases with the higher threshold, the percentage of motifs within the identified regions out of all those in the genome increases with the higher thresholds. By restricting hits to solely those regions identified by both muscle-enrichment techniques at a 2-fold ChIP ratio over background, 2.7% of the CACGTG sites in the genome are found. Given that the regions represent 0.465% of the genome, there is a nearly 6-fold enrichment of the motif over the surrounding genome. The TCTGCG motif shows nearly 3-fold enrichment, the CAACTG motif shows 2.5-fold enrichment, and the GAGACGCA motif shows 3-fold enrichment. However, the TCTCGCAA motif is actually only half as likely to appear in the regions as the rest of the genome. When the regions are restricted further to regions identified with a 3-fold ChIP-ratio over background, the numbers do not change uniformly. Though only 0.07% of the genome falls within these regions, 0.6% of the CACGTG motifs do. This is over 8-fold enrichment over background. The CAACTG motif also increases to 4.75-fold over background. However, the other motifs actually fall in their enrichment. Therefore, the strength of the ChIP-signal correlates with what motifs are found there. Higher ChIP

signals are associated with more E-boxes while lower but still detectable ChIP signals are associated with the accessory motifs.

Dependence of expression on HLH-1 binding

We next asked how transcription factor binding site locations relate to the expression level of genes. The overlap between the data sets will give an idea of how the binding of HLH-1 may or may not affect expression levels in muscle and non-muscle tissue, as well as regarding the impact of the mutation on gene expression levels, both direct and indirect. It is likely that numerous genes that do not have HLH-1 binding sites but that are affected in the mutants are either downregulated due to an indirect regulation from *hlh-1* or upregulated due to regulation by genes upregulated in the absence of *hlh-1*. Genes that are both downregulated in the mutant and have nearby HLH-1 binding sites include *dhp-2*, *hlh-1*, *sup-12*, and *let-2* (Table 7). Some genes, such as *rnt-1*, have binding sites and since they are transcription factors, may have a significant impact on other genes that they in turn regulate. Genes that are upregulated in the mutant and have HLH-1 binding sites include *pgp-10*, in which the binding site falls within the center of the coding region, *rsd-3*, and *tra-4* (There is no correlation between the location of the HLH-1 binding site and whether the gene is upregulated or downregulated in the absence of *hlh-1*). Genes that are downregulated in the mutant but lack any HLH-1 binding site include *bir-2* and *rpl-4*. Interestingly, *unc-120* has no HLH-1 binding site and is not dependent on HLH-1 expression, despite results from earlier time points in other studies that suggest dependence {Yanai, 2008 #1}. There are three other genes apart from *unc-120* that have been identified as a ‘gold standard’ of muscle genes {Fox, 2007 #38} that lack any HLH-1 binding: *tnt-2*, *tmi-3*, and *frm-5*. Though little is known about *frm-5*, both

tnt-2 and *tnt-3* are troponins that are broadly expressed in different types of muscle, both BWM and NSM {Hunt-Newbury, 2007 #62;Ruksana, 2005 #105}. Therefore they may be regulated by *unc-120* or some other pan-muscle transcription factor and not *hlh-1*. Two of the most conspicuous genes upregulated in the mutant, *mls-1* and *hlh-8*, both lack HLH-1 binding sites identified via ChIP-seq anywhere within 20 kb. There remains the possibility that HLH-1 binds to locations not identified with the antibody in anti-HLH-1 ChIP. There are numerous genes with increased expression in wild-type muscle that do not have nearby HLH-1 binding sites identified in our experiments, such as *rpl-2* and *unc-45*.

Correlation of expression with HLH-1 binding motifs

The strength of the anti-body also weakly correlates with the strength of expression. When the binding site is within 500 bp of the start of the gene body, in the 500 bp proximal region, there is a positive correlation between the strength of the binding site and the expression of muscle-enriched genes (Figure 6A). As the signal goes up, when measured by RPKM to normalize for the width of the peak, the expression level tends to also increase. No such correlation is seen with genes not enriched in muscle. Likewise, no correlation is seen when the binding sites occur in other regions, possibly due to a variety of distances and conditions required for regulation (Figure 6B).

DISCUSSION

This study identified both the impact of a lethal mutation on the expression level of embryonic nematode genes and the importance of a transcription factor for proper

function. Muscle differentiation is controlled by at least three transcription factors, of which *hll-1* is the most influential and whose mutation has the most severe impact.

Through the use of RNAi we were able to modify the cell fates of embryonic cells in order to increase the quantity of muscle in the embryo. Because we knocked-down early specification agents to permit muscle specification, artifactual regulatory interactions that can result from engineered overexpression of muscle regulators were avoided. This will avoid both transcription factors driving expression at a greater level than what is physiologically normal, which could lead to transcription factors binding to different target that they would not normally bind at lower, more natural concentrations. Knocking down multiple RNA transcripts can suffer from poor efficiency {Gonczy, 2000 #41}, but our concatenation technique seemed to work well. Also, by driving muscle enrichment in an embryo rather than in a cell culture, our hope was to not activate accessory factors involved in stress response. Nevertheless, there are always consequences to enriching for muscle. At the more extreme end, because the animal is not normal, it will become necrotic sooner than a wild-type embryo. We avoided this side effect by using a relatively early developmental time point prior to hatching. Muscle-enriched embryos might also display unwanted consequences of excess muscle, such as cytokine and signalling imbalances. Despite these muscle isolating caveats, we believe the nuclear enrichment to both be necessary and useful given the notably increased signal.

The basic muscle transcriptome data we gathered corresponds well with existing muscle data. Expected and well-researched muscle genes are present and body wall muscle specific genes, such as the major myosins, calmodulins, and troponins, are in our

muscle-enriched data. We also identified 1915 genes that had not been previously identified as being present in muscle. Similarly, genes enriched in non-muscle correspond well with genes previously identified in other tissues, such as *unc-18* and *glb-7*, which are enriched in neurons. It is possible that some genes specific to muscles from particular lineages are over-represented due to the RNAi knock down. For instance, the C-lineage muscle is overrepresented and any genes that are specific to that lineage may also be overrepresented. However, no muscle is knocked out. By including the *mex-3* RNAi animals we have guaranteed that our view of muscle does not neglect the other lineages. This stage of differentiation is beyond initial specification and into terminal differentiation, so the working assumption is that most lineage-specific muscle variation has passed, as can be seen with the disappearance of *ceh-51* and *hnd-1* expression.

We also identified a set of mutant-affected genes, which corresponds in part with the muscle-affected genes. As expected, the dataset is not identical to the muscle set, as *hlh-1* does not solely control all muscle genes. Clearly, muscle continues to differentiate due to a variety of circumstances and *hlh-1* may have indirect effects on other tissues in which it is not expressed. Many genes that are unaffected are likely driven by additional transcription factors, such as *unc-120*. Surprisingly, many of the important muscle genes that are affected in the mutant are only reduced in expression. Though the *hlh-1* mutation is severe enough to be lethal, its functional absence does not shut down much gene expression. This is most likely due to the continued function of other transcription factors that must work in conjunction with or in the absence of *hlh-1*. Such behavior indicates that multiple cis-regulatory elements drive most muscle genes, or at the very least that HLH-1 binding contributes but is not essential for transcription. The effect of having

overlapping functionality of gene function may play a minor role, but does not appear to be taken advantage of by the network. Of interest is small set of mutant-specific muscle-enriched genes, which may comprise a compensatory reaction of the muscle network to the loss of *hlh-1*.

The upregulation in mutant animals of several transcription factors normally expressed in enteric muscle and the M-lineage post-embryonic sex specific muscle was unexpected. Due to the overlapping targets and transcription factors of these two systems, including such proteins as *myo-3* and *egl-15*, it is not surprising that there is some form of cross-regulation between the systems. Despite being present primarily in non-body wall muscle in healthy wild type animals, *hlh-8* and *mls-1* are both upregulated in the mutant body wall muscle. HND-1 may bind to early *hlh-1* targets before HLH-1 is expressed (Fukushige, 2006 #18). Similarly, though HLH-1 and HLH-8 preferentially bind to different E-boxes {Grove, 2009 #14}, they may bind to sites in front of the same genes and primarily serve the same function.

There is prior evidence for a relationship between *hlh-1* and *hlh-8* in worm muscle development. Although the majority of BWM comes from the embryonic lineages, the post-embryonic M-lineage reveals an underlying relationship. A lack of *hlh-1* causes a number of M-derived body wall muscles to become sex specific muscles {Harfe, 1998 #10, Amin, 2007 #15}. It is possible *hlh-1* deficient embryos exhibit a similar transition. Regulators of *hlh-8* in the NSM such as *unc-62*, *ceh-20*, *lin-39*, and *mab-5*, are present in both wild type and *hlh-1* mutant animals, implying that activation of *hlh-8* and *mls-1* is plausible with the existing architecture {Harfe, 1998 #11}. Somewhat reciprocally, *hlh-8* mutants have an unstable and sometimes higher number of

BWM cells while their sex specific muscles disappear {Corsi, 2000 #48, Corsi, 2002 #44}. This may indicate that there is a balance between the expression of HLH-1 and HLH-8. The proteins may indirectly cross-inhibit each other so that if one is lost, the other turns on. Not all transcription factors normally found in the M-lineage were enriched in our *hlh-1* mutant muscle, including specifically *ceh-24* and *fozi-1*, indicating that there is not a complete fate transformation of the entire tissue. The lack of fate transformation is analogous to the fact that if all myogenic factors are missing, muscle will still not form epithelial tissue (Fukushige, 2006 #18) even though in the absence of *elt-1*, epithelial tissue will form muscle {Spieth, 1991 #49, Michaux, 2001 #27}. The sustained expression of many body wall muscle specific genes in the mutant supports the view that the mutant muscle is mainly body wall muscle in type. Likewise, there is not a strong synthetic PAT phenotype between *hlh-1* and *hlh-8* or *mls-1*, indicating that there is no fate transformation to an *hlh-8* and *mls-1* dependent tissue. Instead it is likely that in the absence of *hlh-1*, *hlh-8* and *mls-1* are indirectly activated and compensate for some of the transcriptional regulation of some muscle genes.

A possible mechanism of *hlh-8* activation in the mutant can be postulated from our results, although a complete dissection of the pathway is beyond the scope of this study. This method of *hlh-8* activation may also play a role in the NSM. From the synthetic PAT phenotype analysis, *ceh-20*, *lin-39*, and *grh-1* were among genes identified as strong genetic interactors with *hlh-1*. By independent criteria, each of these is also a candidate to help activate *hlh-8*. Thus, *grh-1* regulates *mab-5* {Venkatesan, 2003 #51} and, based on interactions known in *Drosophila*, interacts with *egl-15* {Zhong, 2006 #33}. EGL-15/FGFR is necessary for proper sex myoblast migration {Stern, 1991 #66}.

The splicing variant EGL-15a is preferentially expressed in sex myoblasts and vulval muscle. It is downregulated by SUP-12, which destroys EGL-15a but not EGL-15b, the splicing variant primarily expressed in body wall muscle {Kuroyanagi, 2007 #104}. We found that *sup-12* expression depends on HLH-1 binding. We also found that the EGL-15a splicing isoform is upregulated in the mutants. Therefore, in healthy body wall muscle, HLH-1 drives SUP-12 to downregulate EGL-15a, while in muscle without HLH-1, SUP-12 is not expressed and EGL-15a increases, the variant found in NSM.

Both *mab-5* and *lin-39* are known to interact with *ceh-20* and *unc-62* to either activate or repress genes, depending on which Hox gene is dominant {Liu, 2006 #50, Jiang, 2009 #25, Potts, 2009 #26}. Since *mab-5* is implicated in the proper formation of *hlh-8* dependent cells {Kenyon, 1986 #43}, it is possible that for muscle tissue active MAB-5 promotes sex muscle development while body wall muscle depends on LIN-39. In *mab-5* mutants, some body wall muscle ends up as SM cells, indicating that *mab-5* may instead be necessary for body wall muscle development {Harfe, 1998 #10}. The interaction between these two Hox genes may be more subtle or complex than requiring either gene to be active or inactive. Much of the Hox/Pbs/Meis complex function is controlled by nuclear localization rather than transcriptional regulation, explaining the consistent expression between the mutant and wild type animals {Jiang, 2009 #25; Potts, 2009 #26; Liu, 2006 #50}. It is possible that in the absence of *hlh-1*, the upregulation of *grh-1* in combination with extra EGL-15A leads to the activation of either the *lin-39/ceh-20/unc-62* or *mab-5/ceh-20/unc-62* complex, thus driving the expression of *hlh-8*. HLH-1 binding to enhancers around *mab-5* and *unc-62* suggests that *hlh-1* may serve as the counter to *grh-1*, though it is not known what might activate *grh-1* in the absence of *hlh-1*.

or how it is repressed in wild type animals. Because *hllh-8* does not present a synthetic phenotype with *hllh-1*, it is likely that there are multiple targets of this complex, possibly including *mls-1*.

hllh-1 and *hllh-8* are similar proteins and they play similar roles in muscle formation in their respective tissues. Their homologs in other species also play similar roles, suggesting that there is an ancestral basis for this functional overlap. In *Drosophila*, the *hllh-8* homolog *twist* plays a central role in muscle formation while the *hllh-1* homolog *nautilus* plays a non-essential role in muscle differentiation {Balagopalan, 2001 #52}. In vertebrates, Twist plays a role in a subset of muscle formation {Castanon, 2002 #53} while the MRF genes, homologs to *hllh-1*, play a central role in skeletal muscle formation. This is bolstered by the presence of *mls-1/Tbx1* being upregulated, as in vertebrates it is also involved in skeletal muscle development {Chieffo, 1997 #54}. The overlapping functions of these genes in muscle formation suggest that they have a common evolutionary history. Thus the activation of one in the absence of the other may be the result of an ancient evolutionary network or an accident of similar systems that use similar components.

By identifying genome-wide endogenous HLH-1 binding sites we have a more complete understanding of both transcription factor behavior in *C. elegans* and the functional role of *hllh-1*. By using an existing antibody with an untagged HLH-1 epitope, we were able to observe HLH-1 binding in unmodified animals. This method has dual advantages: no competition between tagged and untagged HLH-1 and no tampering with expression levels. Tagged versus untagged ChIP-seq has only been performed against RNA Pol II {Zhong, 2010 #207} and further comparisons of transcription factor tagging

will be necessary in the future to determine its merits and consequences. As one of the principal objectives is to understand how gene expression changes when the transcription factor is mutated, not tampering with expression levels is critical. This complements the increase in body wall muscle quantity without directly modifying the myogenic regulatory factors. One possible limitation is the shielding of HLH-1 if other protein complexes are also bound, though similar problems can arise if a tag interferes with complex formation. A major downside is the limited sensitivity of the antibody. As some antibodies are naturally more sensitive than others, it is possible that some of the signal is lost at sites of reduced binding. Nonetheless, we were able to map close to a thousand binding sites throughout the genome. The genes were observed to be concentrated in the proximal upstream region of genes and intergenically, with much smaller representation within gene bodies. There was some binding within introns, but it was by no means the primary binding location. Binding within exons is disenriched. Only one instance was observed of a binding site within a gene body where the gene was expressed more strongly in the mutant. This suggests that if *hlh-1* has any repressor functions – which are most likely limited due to the small number of genes upregulated in the mutant with nearby binding sites – it does not primarily function by binding to the gene body to prevent transcription. The majority of the binding, however, appears to have no impact on nearby genes. Some of these genes are likely affected by *hlh-1* and simply buffered by other transcription factors in the mutant. However, many binding sites are likely ineffective and serve no biological function, as has been seen in other organisms (Cao et al., 2010).

From the anti-HLH-1 selected regions we were able to reconstruct the two E-box binding sites of HLH-1 as well as other associated sites. The E-boxes had been previously identified both *in vitro* (CAGCTG, Grove, 2009 #14) and at certain sites (CAACTG, Fukushige, 2005 #19). Our data has served as an *in situ* genome-wide confirmation of these prior predictions. Stronger binding correlates with strong E-boxes. It is possible that the weaker the E-box, the more additional factors are needed to improve transcription factor binding. The E-box, being generally centered on the experimentally observed HLH-1 binding site, is almost certainly the actual binding motif. The TCTGCG motif is also well-centered and may serve as an alternate, weaker binding site. Less centered are the GAGACGCA and TCTCGCAA motifs, which bear resemblance to muscle-associated motifs {Guhathakurta, 2002 #32; GuhaThakurta, 2004 #35} and may be binding sites for accessory factors that can recruit HLH-1 to regions that lack a suitable binding site or have a weak binding site.

The effect of *hll-1* mutation on developing embryonic muscle is severe but intriguingly addressed by the transcriptional machinery. While some buffering may arise from overlapping gene functions, this does not appear to be the primary compensatory mechanism. Though expression levels of target genes fall, few genes are shut off as other regulators keep them transcribed. And while many factors decrease their impact, some other transcription factors are brought into play to assist in muscle formation.

MATERIALS AND METHODS

General methods and strains. We obtained *Caenorhabditis elegans* from the CGC strain collection and cultured them using methods standard for *C. elegans* {Sulston, #56}. Strains used included PD4605 (*hlh-1(cc561)*) and N2. Worms were grown up at 20°C (15°C for the temperature sensitive mutants) on HT115 bacteria.

RNAi feeding. Bacteria from the Ahringer Lab RNAi Library were utilized for the HT115 empty vector (no RNAi) feeding and for *mex-3* RNAi feeding. The inserted sequences from *mex-3* and *skn-1* were inserted in the *elt-1* vector to generate the triple RNAi feeding vector using restriction enzymes. 8cm NGM special plates with IPTG and carboxy-penicillin were seeded with the RNAi bacteria, grown in LB and concentrated to 20% w/v prior to plating. Plates were dried and the worms were added and grown until the L4 stage, at which point the temperature was increased to 25°C until the animals were gravid and egg-laying had begun.

Harvesting worms. Gravid adults were washed off plates and bleached to obtain eggs. The eggs were then shaken in S-complete medium at a density of 5 embryos/ μ L for 400 minutes. The embryos were spun down and prepared for the desired set of observations.

Chromatin preparation and ChIP. For DNA-based sequencing we used a modified version of the Farnham ChIP-seq protocol {Weinmann, 2002 #208}. Embryos were suspended in 2% formaldehyde, freeze cracked on dry ice 5 times, allowed to fix for 30 minutes, and then quenched with Tris-HCl for 5 minutes. The embryos were washing in Tris, Farnham Lysis Buffer, and RIPA buffer. They were then sonicated (Misonex

model at power output 3.5) with a microtip for 15 30-second pulses with 1 minute cooling intervals. 10% of the sample was set aside for purification without antibody addition. The antibody was added to the chromatin prep and allowed to mix for 16 hours at 4°C. 200 µL of magnetic beads (Invitrogen Dynabeads M-280 Sheep anti-Rabbit IgG) were then added for 4 hours to extract the antibody. The addition of beads was repeated 3 times and all beads were pooled. The beads were washed and the complexes eluted and purified with a phenol-chloroform-isoamyl alcohol precipitation. The DNA was quantified with a fluorometer (Invitrogen Qubit Fluorometer).

mRNA purification. For RNA-based sequencing, embryos were flash frozen in trizol (Sigma) and freeze-cracked on dry ice 5 times. The embryos were then passed through a 21 G needle ten times, followed by a 25 G needle an additional 10 times to help shear the eggshell. The RNA was then purified with a standard Trizol-chloroform precipitation. A dT purification was then performed with magnet beads (Invitrogen Dynabeads Oligo-dT).

Library making and sequencing. The standard Illumina library-making protocol for single amplification was used, including end repair, adaptor ligation, gel purification, and PCR amplification. The Illumina protocol was also followed for flowcell generation and sequencer running.

RNAi feeding for synthetic lethal screening. Bacteria from the OpenBioSystems RNAi library and the Ahringer RNAi library were used for RNAi feeding of L4 animals for 36 hours at 25°C. Adults were then transferred to fresh plates for egg-laying for 4

hours at 25°C. Adults were removed and embryos were allowed to develop for 18-24 hours prior to scoring.

Scoring. Embryos were scored for developmental progression using a dissecting microscope. The stage of developmental arrest in embryonic lethal worms was noted as during the two-fold stage (PAT) or otherwise.

Data Analysis. For all data analysis Wormbase release WS190 was used. Read mapping was performed with Bowtie and preliminary data analysis was performed with ERANGE {Pepke, 2009 #31}. Extended data analyses were performed using original code in Python. The requirement for a gene to be categorized as having muscle-specific expression is that the muscle-enriched mean expression minus the standard deviation must be greater than the muscle-normal mean expression plus its standard deviation. Genes associated with stress response (such as heat shock genes) were checked for expression to guarantee there was no sign of damage or stress to the embryos.

ACKNOWLEDGEMENTS

Some nematode strains used in this work were provided by the Caenorhabditis Genetics Center, which is funded by the NIH National Center for Research Resources (NCRR). We would like to thank M. Krause for the anti-HLH-1 antibody, K. Fisher for her assistance in generating density graphs, H. Amrhein for assistance in managing genomic databases, and J. DeModena for assistance in obtaining reagents. This work was supported by HHMI, for which PWS is an investigator.

TABLES:**Table 1: Sample of genes upregulated in muscle enriched animals**

1394 genes are enriched in muscle-rich animals. A subset of these genes, including examples of well-documented muscle structural genes {Fox, 2007 #38}, is given.

Gene	Gene Description
act-2	actin
act-4	actin
deb-1	vinculin (dense bodies)
dhp-2	dihydropyrimidinase
dim-1	immunoglobulin-repeat (myofilament anchoring)
egl-15	FGF-like receptor tyrosine kinase
egl-19	alpha subunit of mammalian L-type calcium ion channel
egl-20	WNT
emb-9	basement membrane collagen
epi-1	laminin alpha chain
let-2	alpha-2 type IV collagen
lev-11	tropomyosin
lin-1	ETS transcription factor
lin-2	membrane associated guanylate kinase
lin-39	sex combs reduced/Hox5
mup-2	troponin T
myo-3	myosin heavy chain A

pat-3	beta-integrin subunit
tmd-2	tropomodulin
tni-1	troponin
tnt-3	troponin T
unc-112	Mitogen inducible gene- (dense bodies and M lines)
unc-116	kinesin-1 heavy chain
unc-15	paramyosin
unc-23	chaperone
unc-44	ankyrin-like protein
unc-45	chaperone
unc-52	perlecan
unc-53	NAV1/2/3
unc-54	myosin class II heavy chain
unc-68	ryanodine receptor
unc-70	beta-spectrin
unc-73	guanine nucleotide exchange factor
unc-83	transmembrane protein
unc-89	protein kinase (A bands)
unc-94	unknown (thin filaments)
unc-95	paxillin-related (thick and thin filaments)
unc-96	unknown (thick filaments)

Table 2: Genes upregulated in the *hlh-1* mutant muscle that are known to be expressed in NSM

Of the 307 genes that are upregulated specifically in the mutant muscle but are not enriched in wild type muscle, only 96 have described expression patterns. Of those, a full 27% have been observed in at least a subset of NSM. The function of these genes is described in this table.

Gene	Description
B0336.3	RNA recognition
ags-3	G protein signaling
arr-1	beta-arrestin
C03H5.2	UDP transporter
ced-1	lipoprotein receptor
cts-1	citrate synthase
dpy-23	adaptin
dsc-1	defecation suppressor
egl-20	WNT, signalling protein
exp-1	GABA receptor
F47B7.2	sulphydril oxidase
H28O16.1	ATP synthase
hlh-8	TWIST, transcription factor
mls-1	TBX1, transcription factor

mrp-2	Multi-drug resistance protein
mua-6	intermediate filament
mup-4	muscle junctions
nlp-13	neuropeptide
nmy-1	non-muscle myosin
ppk-3	PIP kinase
rom-1	rhomboid related
shc-1	signaling (src, jnk, insulin)
snb-1	synaptic vesicle
trs-1	tRNA synthetase
uvt-3	pantothenate kinase
ZK112.3	unknown

Table 3: Synthetic PAT Scoring

Transcription factors were screen for synthetic paralysis at the two-fold stage (PAT) using RNAi feeding in *hlh-1(cc561)* mutant animals. Several genes gave significant increases in the phenotype in the mutant background, including the genes *lin-*

39, *grh-1*, and *ceh-20*. Shown are the percentage of PAT phenotype seen in the screen and the significance.

RNAi feeding	Wild type (N2)	<i>hlh-1(cc561)</i>
No RNAi (HT115)	0% (0/411)	5% (4/88)
<i>ceh-20</i> RNAi	0% (0/161)	31% (57/121)
<i>lin-39</i> RNAi	1% (1/112)	27% (27/101)
<i>grh-1</i> RNAi	0% (0/250)	25% (23/93)
<i>nhr-63</i> RNAi	1% (2/141)	27% (27/99)

Table 4: Location of peaks relating to gene bodies

Peaks are located in various locations surrounding genes. They are not necessarily functional from each of these locations, as gene models overlap and it is not always upon which gene the transcription factor is acting. In the table the number of motifs and frequency of the motifs in different regions of the genome are shown. The numbers add up to more than 100% due to overlapping gene bodies and the regions not being mutually exclusive. The Gene Body refers to the exons and the introns, the Exons includes the CDS and the 5' and 3' UTRs, the CDS refers to the coding sequence (translated exons only), and both the exons and introns are counted more than the Gene Body due to different isoforms being counted more than once. As can be seen, the greatest enrichment in binding is in the 500 bp proximal promoter region, followed by the 5'UTR, and finally the upstream intergenic region. Other regions of the gene bodies are depleted for binding, though some is still present.

Region	Bases counted	Number of peaks	% of 9447 peaks	Fold enrichment
5000 bp Upstream	119250000	16323	173%	0.1

500 bp Upstream	11925000	13121	139%	0.9
Gene Body	62380610	6897	73%	1.2
Exons	41536270	6597	70%	1.7
Introns	51963032	899	9.5%	0.2
CDS	61363568	6609	70%	1.1
5' UTR	694154	45	0.48%	0.7
3' UTR	2540801	36	45%	0.2

Table 5: Genes with HLH-1 binding nearby

Many genes throughout the genome have HLH-1 binding within the gene body or in the 5' 5000 base pairs. Both HLH-1 dependent genes and muscle-enriched genes are more likely to have binding than the rest of the genome. Genes that are upregulated only in mutant muscle are actually less likely to have HLH-1 binding than the background level.

	Number of genes with binding	Total number of genes	Percent of genes with binding
HLH-1 dependent genes	584	1070	54.6%
HLH-1 dependent genes that are enriched in muscle	120	216	55.6%
Other HLH-1 dependent genes	464	854	54.3%
Genes enriched in muscle	1169	2175	53.7%
Genes enriched in muscle that are not dependent on HLH-1	1049	1959	53.5%
Genes enriched in <i>hll-1</i> mutants	224	415	54.0%
Genes enriched only in <i>hll-1</i> mutant muscle	119	308	38.6%
All genes with no expression dynamics	6501	13477	48.2%

All genes not enriched in muscle	8350	18173	45.9%
All genes not dependent on HLH-1	8935	19278	46.3%

Table 6: Regions in which signal is found compared to signal intensity

As shown here, there is a strong correlation with the peak strength and the identifiable motifs present. All but the strongest ChIP signals disappear with 3-fold enrichment of peaks over background. Therefore, motifs where the fold enrichment of motifs found within peaks compared to background level (number found throughout the genome) increases from 2-fold to 3-fold, that indicates that those motifs are associated with stronger signals. The E-boxes, CAGCTG and CAACTG, correlate well with very high ChIP signals. If the fold enrichment decreases with higher peaks, then those motifs are associated with weaker ChIP signals, which is true for the non-E-box motifs. Also shown is the relation between strict motif-finding with a 95% threshold and looser motif-finding with a 85% threshold. In all cases, a higher threshold leads to greater enrichment, indicating that the motifs are more likely to be found in the ChIP peaks than degenerate motifs.

Motif	Threshold	# motifs in genome	Within 2-fold peaks	Fold enrichment over genome representation	Within 3-fold peaks	Fold enrichment over genome representation
CAGCTGTT	85%	77845	925	1.4	297	1.9
	95%	3370	150	5.8	55	8.4
CTCTGCGT	85%	38054	596	1.1	111	1.0
	95%	2795	94	2.8	18	1.5
CAACTGTT	85%	137496	749	0.53	203	0.81
	95%	5369	107	2.6	41	4.8
GAGACGCA	85%	48039	753	0.98	128	0.74
	95%	5369	258	3.0	40	1.6

TCTCGCAA	85%	80846	431	0.32	89	0.35
	95%	2662	26	0.48	6	0.53

Table 7: Correlation of HLH-1 binding and decreased mutant gene expression

Numerous genes lose a significant amount of expression in the muscle-enriched animals in the absence of HLH-1. Some examples are included below.

Gene	Function
alh-8	aldehyde dehydrogenase
cic-1	claudin
clec-92	C-type lectin
cyn-10	cyclophylin
dhp-2	dihydropyrimidinase
ech-2	enoyl-coA hydratase
etr-1	RNA binding
fbxb-37	f-box b
fem-3	feminization
ife-4	initiation factor
lact-9	beta-lactamase
let-2	muscle collagen
let-756	FGF ligand
lev-11	tropomyosin
mig-17	metalloprotease
mys-1	histone acetyltransferase
ndx-4	NUDIX hydrolase
npp-20	nuclear pore complex
ost-1	basement membrane osteonectin
pfd-5	molecular chaperone
pup-3	polyU polymerase
rnt-1	RUNX transcription factor
rpl-32	ribosome

rps-30	ribosome
rps-4	ribosome
rps-8	ribosome
rsp-6	ribosome
sft-1	Surf1
sfxn-5	mitochondrial iron transporter
sup-12	muscle specific RNA binding
syg-1	transmembrane immunoglobulin
syg-2	transmembrane immunoglobulin
tag-165	methionine synthase reductase
tnt-3	troponin
tsp-11	integral membrane tetraspanin
tsp-17	integral membrane tetraspanin
ttr-16	transthyretin
twk-31	potassium channel
ubc-19	ubiquitin conjugating enzyme
ugt-24	UDP glucuronosyl transferase
zmp-1	zinc metalloprotease

FIGURES:

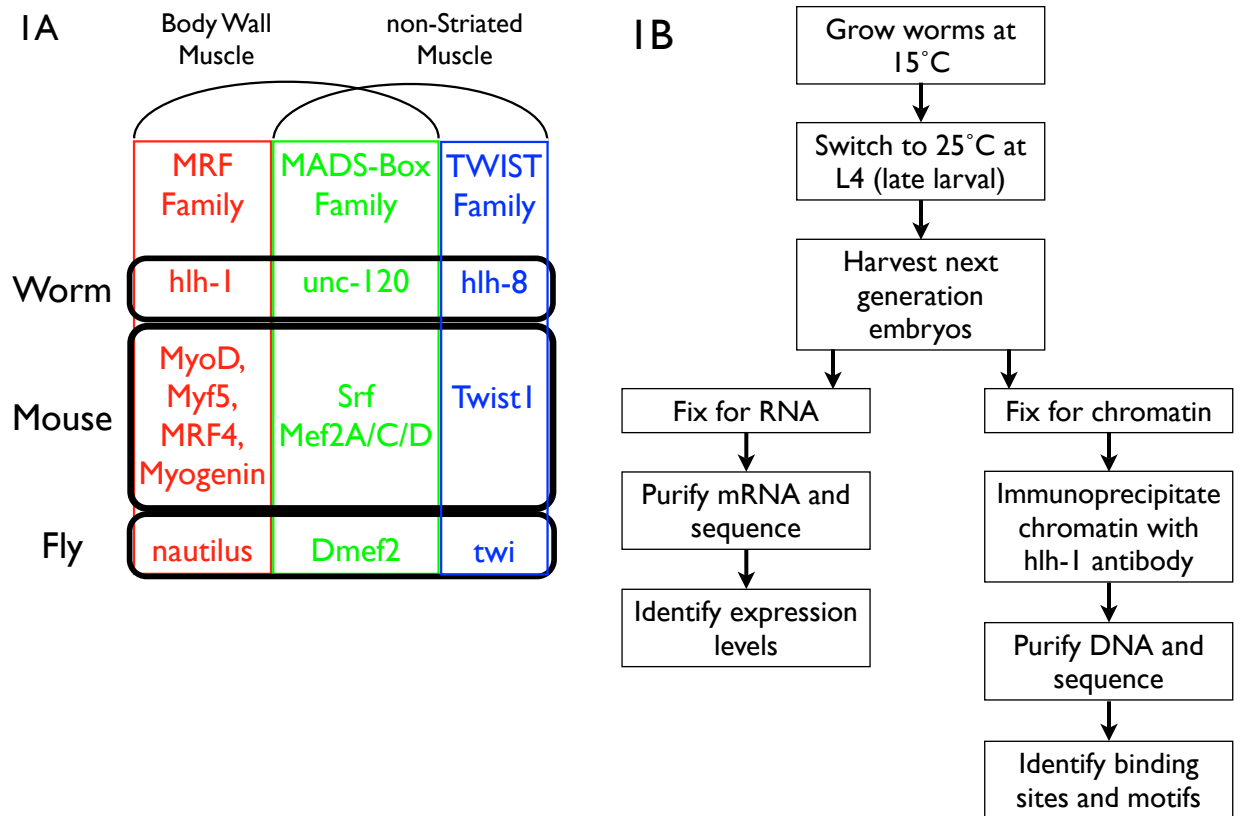


Figure 1: Experimental flow and muscle differentiation network

(A) Families of transcription factors involved in muscle differentiation observed across phyla with their worm expression tissues highlighted. (B) The experimental rationale is shown.

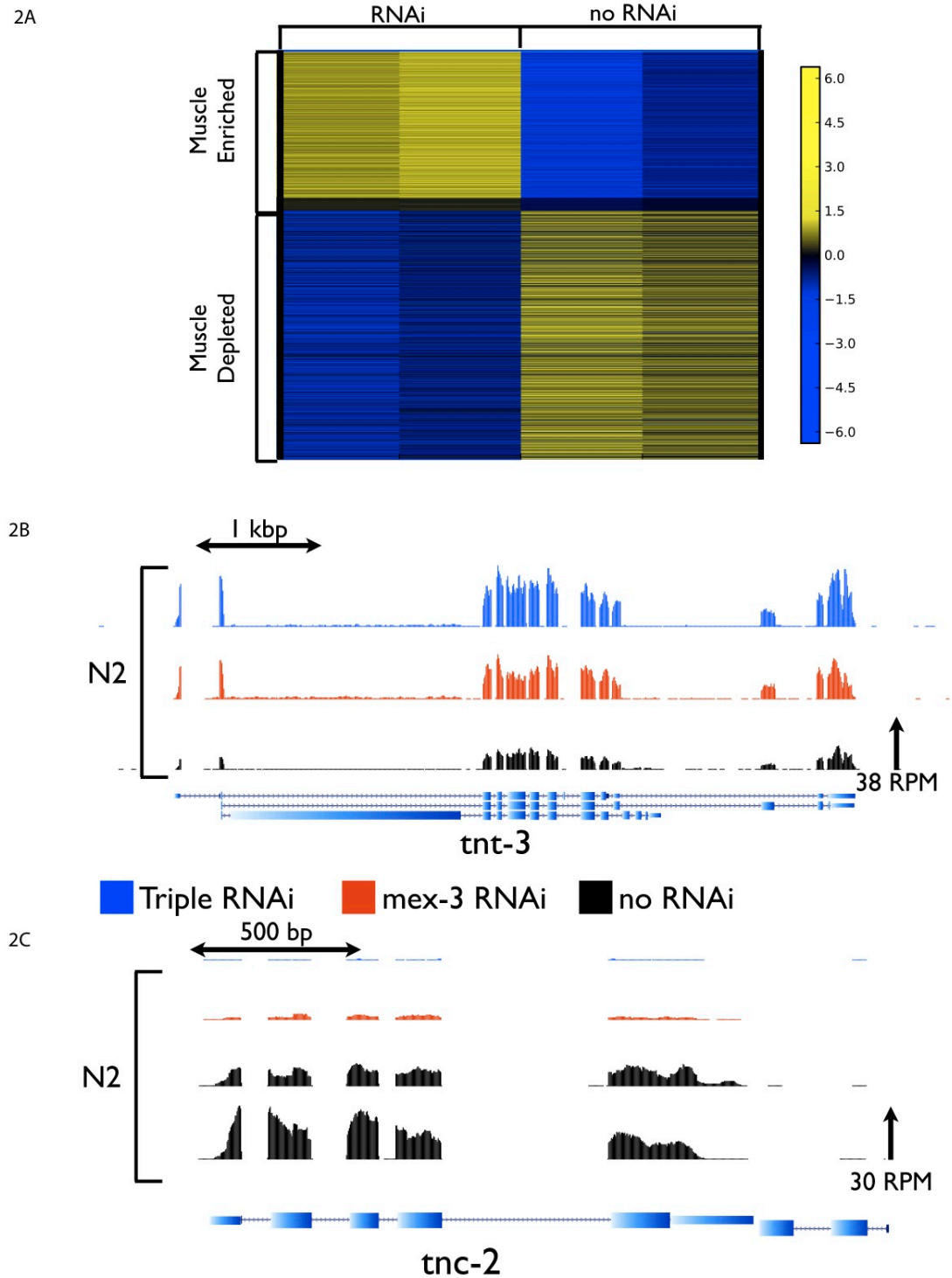
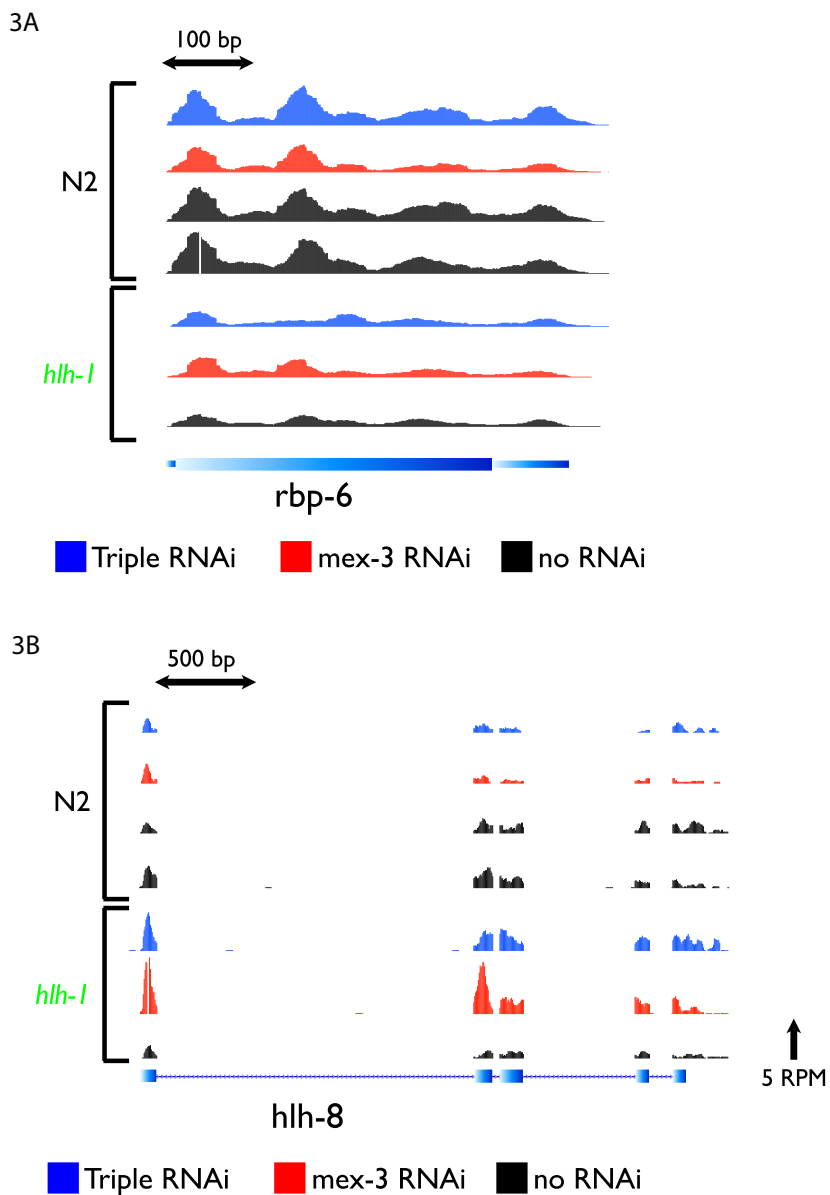


Figure 2: The impact of RNAi-based muscle enrichment on gene expression levels

(A) An RPKM heat map of expression levels determined by RNA-seq of poly-dT selected mRNA across normal (no RNAi) and muscle enriched (*mex-3* RNAi and *mex-*

3/elt-1/skn-1 RNAi) conditions. Higher expression is in yellow, lower expression is in blue. (B) The expression of the muscle troponin T, *tnt-3*, is shown, with muscle-normal expression in black and muscle enriched in blue (triple RNAi) and red (*mex-3* RNAi). Since muscle-normal animals still have a significant amount of muscle, expression is still seen. (C) The expression of the non-muscle troponin C, *tnc-2*, is shown. Since the triple RNAi animals (blue) have very little non-muscle tissue, very little expression is seen.



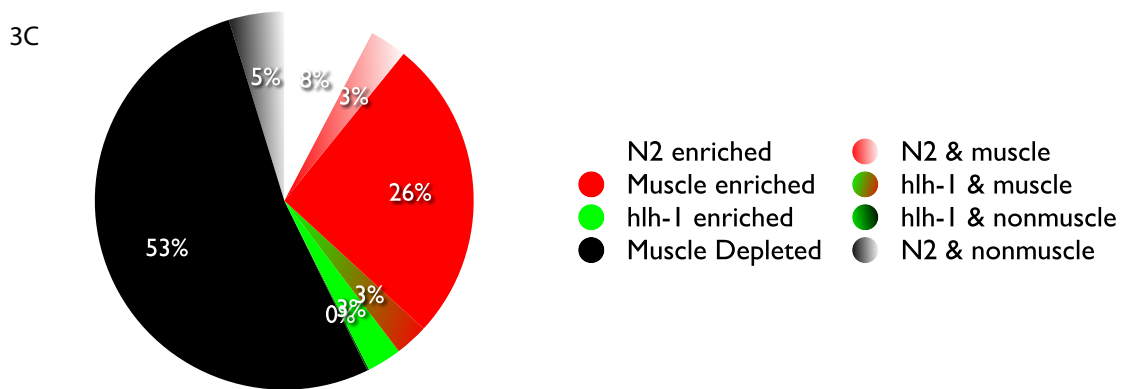


Figure 3: The impact of mutation on gene expression levels

(A) The expression of a gene, *rbp-6*, enriched in wild-type animals (top 4 lanes) across RNAi conditions but with decreased expression in the *hlh-1* mutant (bottom 3 lanes) (B) The expression of the nematode Twist, *hlh-8*, is shown. Its expression is increased primarily in the muscle enriched mutant animals, with the muscle-enriched wild-type animals having less expression than the muscle-normal wild-type (C) The relationship of genes that are enriched in wild-type versus mutant animals and muscle-normal versus muscle-enriched are shown. The relative number of enriched genes for muscle (red), non-muscle (black), wild-type (white), and *hlh-1* mutant (green) are shown, along with genes that are enriched in more than one category.

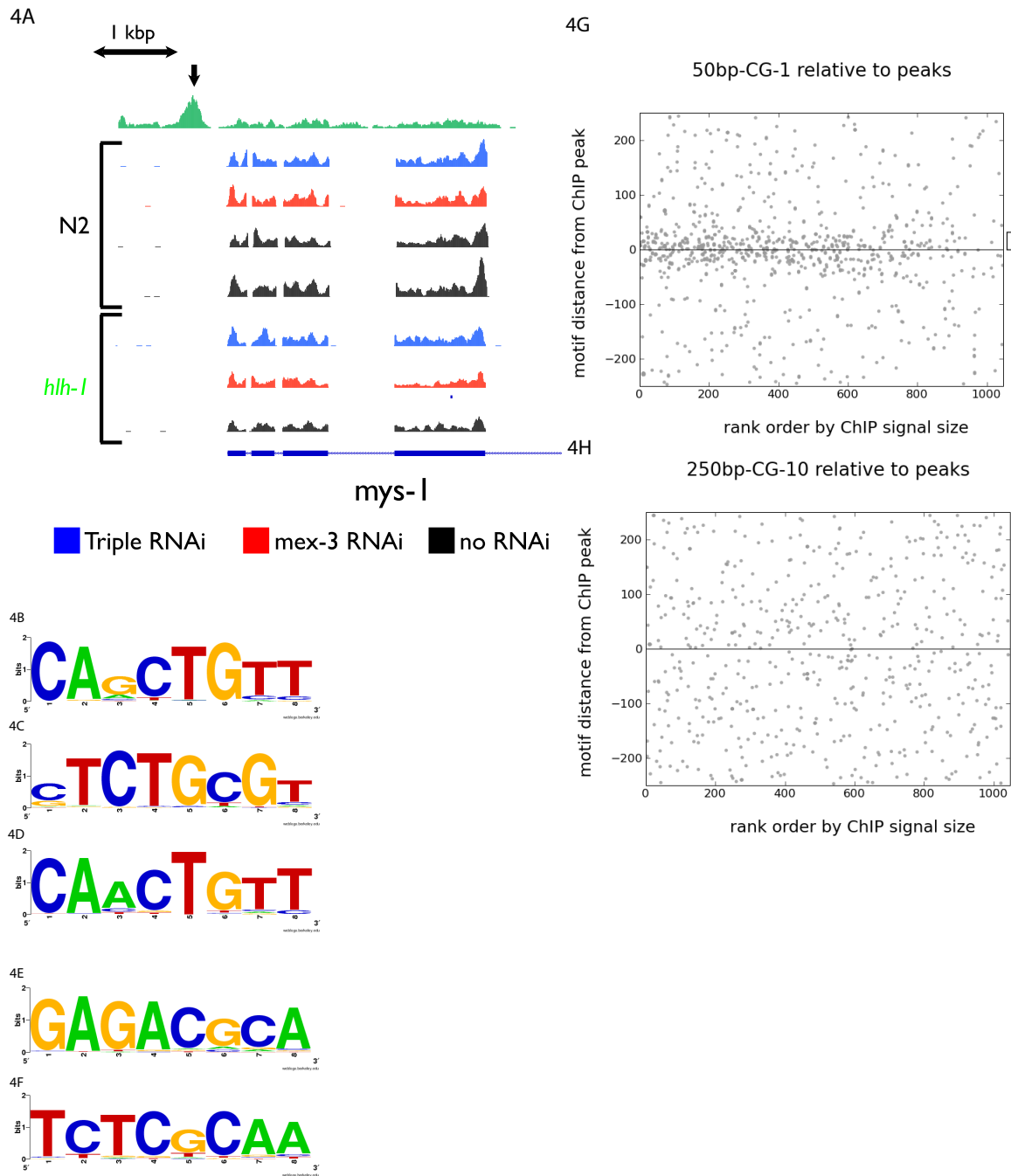


Figure 4: HLH-1 binding results determined from anti-*hlh-1* ChIP-seq

(A) The wiggle-gram is an example of the signal from Anti-*HLH-1* ChIP-seq with a peak at the *mys-1* locus. The relative expression patterns of the wild-type and mutant are also shown. *mys-1* is expressed in the mutant but at much lower levels, indicating its

dependence on HLH-1 binding for expression. (B-F) The weblogo PSFM diagrams for the top non-repeat motifs found. (B) The primary motif identified within a 50-bp radius of the hlh-1 binding site (C) The 8th identified motif at a 50-bp radius. (D) The top motif at a 100-bp radius. (E) The top motif at a 250-bp radius. (F) The tenth motif at a 250-bp radius. (G) Shown here is the relative location of the CACGTG motif compared to the experimentally identified binding site. The motif is clearly centered on the binding site and is tightly bound to the center. (H) By comparison, the GAGACGCA motif shows no centrality or correlation with the binding site. (I) Comparison of peak strength to motifs found.

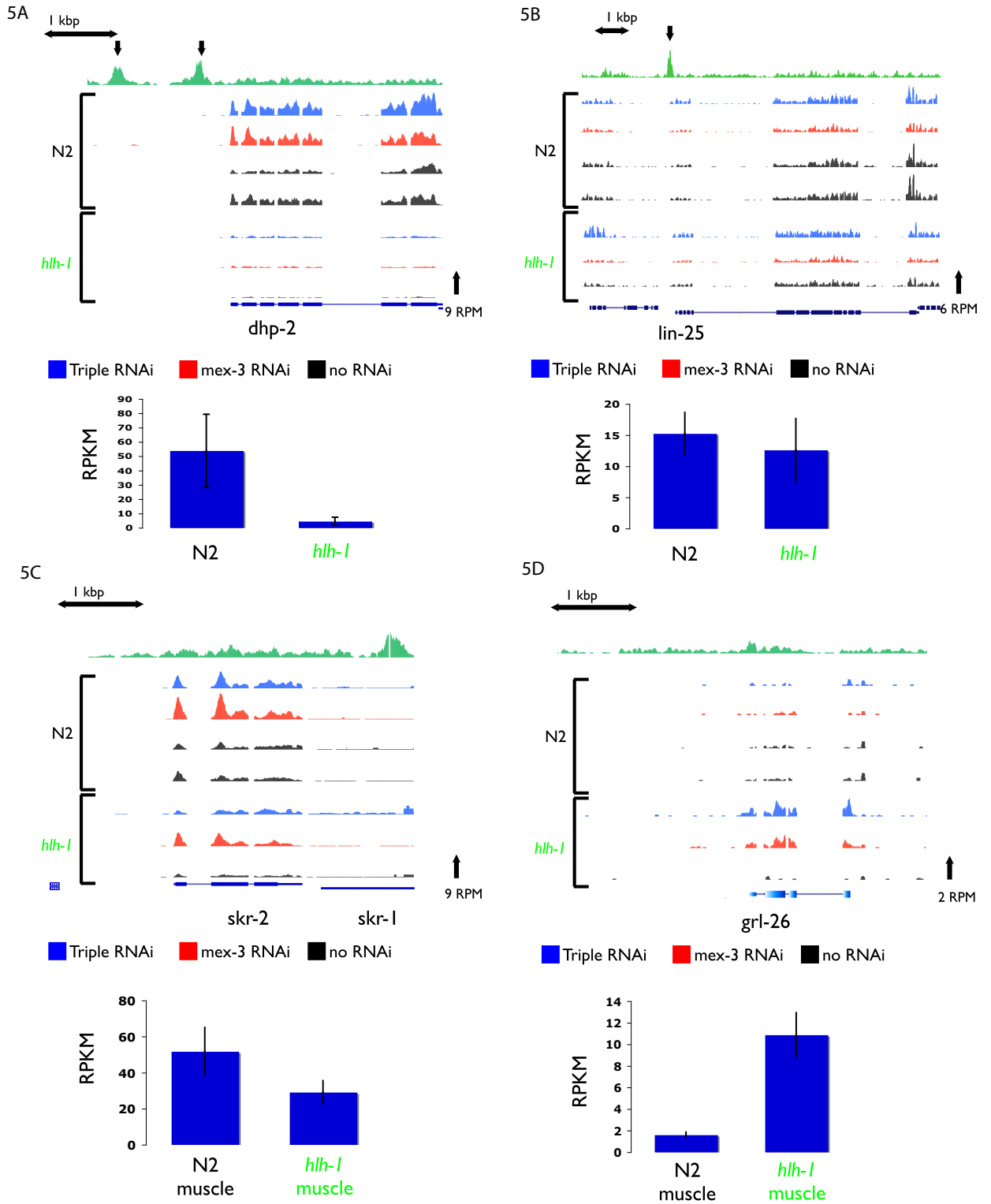


Figure 5: Correlation between expression and binding

The correlation of *hlh-1* binding in the top track with expression in both wild-type and the mutant. (A) *dhp-2*, representing genes with high *hlh-1* binding (the top green track) and a large decrease in expression in the mutant (bottom three tracks) is shown. These genes are likely dependent on *hlh-1* driving expression. The bar graph shows the comparative expression level in RPKM between wild type and the mutant. (B) *lin-25*, with high *hlh-1* binding but with little change in expression in the mutant is shown. These genes may be driven by *hlh-1* but are sufficiently buffered by other transcription factors. The bar graph shows the comparative expression level in RPKM between wild type and the mutant. (C) *skr-2* with no upstream *hlh-1* binding but with a large decrease in expression is shown. Most likely the *hlh-1* peak in the downstream gene *skr-1* is involved in regulating *skr-2*. The bar graph shows the comparative expression level in RPKM between wild type muscle and mutant muscle. (D) *grl-26* with no *hlh-1* binding but with a large increase in expression is shown in mutant muscle. This gene is likely turned on only in the absence of *hlh-1*, probably indirectly rather than repression by *hlh-1*. The bar graph shows the comparative expression level in RPKM between wild type muscle and mutant muscle.

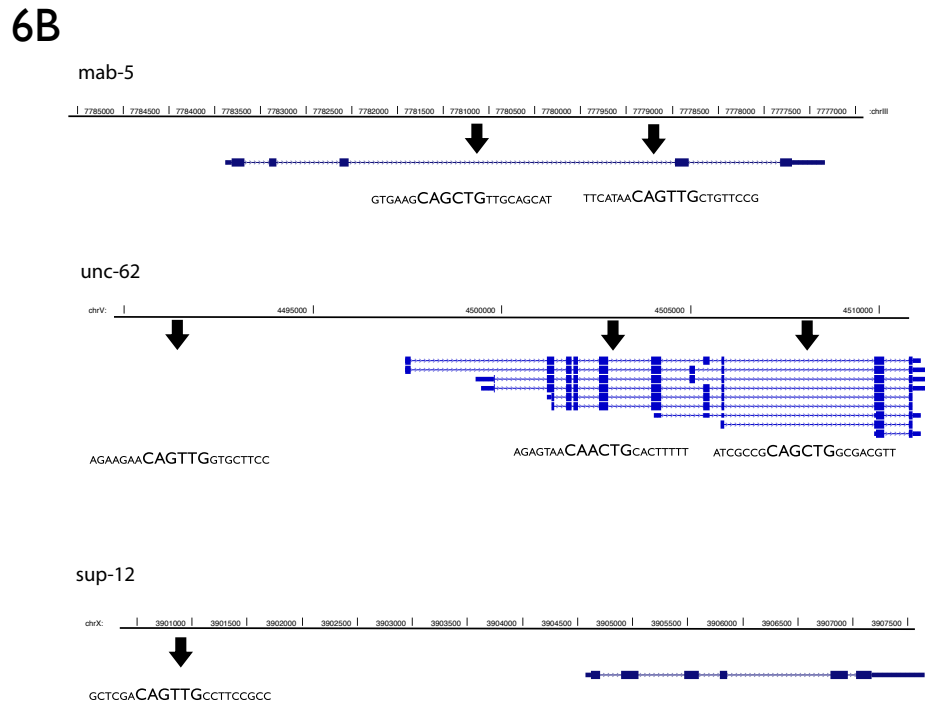
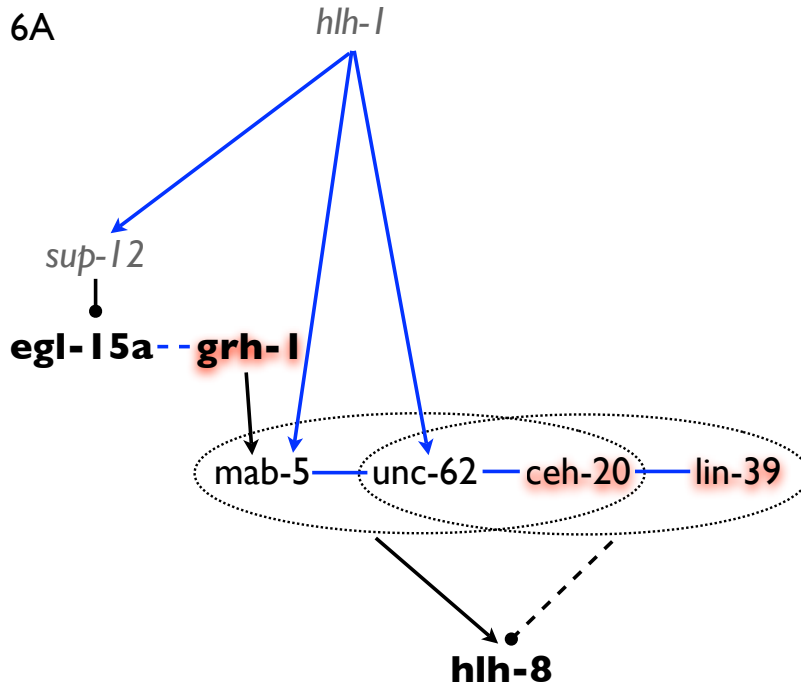


Figure 6: Proposed model of *hlh-8* activation

A number of genes are known to be involved in a potential pathway to activate *hlh-8* based on both existing and new data. (A) In the absence of *hlh-1*, *grh-1* is upregulated (in bold) and may be controlled by *egl-15* (dashed blue line) {Zhong, 2006 #33} to regulate *mab-5* {Venkatesan, 2003 #51}. The splicing variant EGL-15a is inhibited by the mRNA-binding protein *sup-12* {Kuroyanagi, 2007 #104}. *sup-12* is dependent on HLH-1 binding for expression. In turn, MAB-5 competes with LIN-39 to interact with CEH-20 and UNC-62 in some cells to effect target expression or repression (blue lines) {Jiang, 2009 #25;Liu, 2006 #50;Potts, 2009 #26}. A similar action may be occurring here, with *hlh-8* known to be dependent on *ceh-20* and *mab-5* in some cells {Jiang, 2009 #25;Kenyon, 1986 #43;Liu, 2006 #50}. Therefore, in the absence of *hlh-1*, *sup-12* is downregulated – leading to an increase in EGL-15a – and *grh-1* is upregulated. GRH-1 and EGL-15a work together to activate the MAB-5/UNC-62/CEH-20 Hox/Pbx complex to upregulate *hlh-8* (in bold). As our data shows HLH-1 binding near *sup-12*, *mab-5*, and *unc-62* (blue arrows), it may serve to repress *hlh-8* by means of this pathway (dashed arrow). The existence of this pathway is also supported by the appearance of *grh-1*, *ceh-20*, and *lin-39* (shadowed in red) in the synthetic PAT screen as being necessary for muscle formation in the absence of *hlh-1*. (B) HLH-1 binding sites for *mab-5*, *unc-62*, and *sup-12* are shown.

REFERENCES:

Chapter 5: Conclusions

Future directions in developmental regulatory network analysis

The goal of this series of research projects was to expand our knowledge of developmental regulatory network function. By identifying a collection of cis-regulatory elements, transcription factor binding sites, and transcription factors we have expanded our knowledge of network components. From studying the interactions between many of these parts we have been able to more accurately map the structure of regulatory networks. *C. elegans* has proven to be a convenient model organism for transcriptional study on several levels and advances in understanding nematode regulatory networks should translate to other organisms.

To begin expanding the complement of known network components, we first demonstrated the utility of a bioinformatic technique, ungapped evolutionary sequence conservation, for identifying non-coding cis-regulatory elements. By comparing sequences across four closely related *Caenorhabditis* species, including one novel set of sequence, we identified cis regulatory elements at high efficiency. The elements discovered in the Hox cluster were shown to be functionally independent and more complex than a single transcription factor binding motif. We established parameters for identifying regulatory elements within a complex but evolutionarily consistent locus. By testing both regions predicted to be functional and those expected to lack regulatory elements we were able to quantify our technique's efficiency. We predict that equivalent parameters should yield similar rates of success in other loci with comparable degrees of evolutionary age. This was demonstrated by our ability to recapitulate known enhancers in well-dissected promoters. We were even able to identify an ancient regulatory element that was conserved between vertebrates and nematodes. Past research has shown that

some elements taken from one organism can function in an organism from another phyla, but our data points to functional elements in two phyla that share a common origin. While very old and well-conserved elements are ideal to identify with this comparative sequence analysis, further research will be needed to identify the limitations of this technique at other loci. Rapidly diverging gene functions are connected to rapidly evolving regulation. A genome-wide, large scale sampling study is needed to complement the findings of our initial pilot study. The Hox cluster, being very highly conserved, is in many ways an ideal case. Future regulatory element analyses should include rapidly evolving genes in addition to the very highly expressed, highly conserved genes. Rapidly evolving cis-regulatory elements controlling neuron signalling may require a very different set of comparison parameters, both at the level of what genomes to compare and what thresholds and windows to use.

We investigated the relationship between different transcription factors in the muscle differentiation network to increase the catalogue of factors pertinent to muscle development. We used an RNAi synthetic PAT phenotype screen to identify a number of factors that range from necessary for healthy muscle development to interacting with the network only under duress. These transcription factors can be grouped into four broad categories. The first category consists of transcription factors that normally participate in muscle specification and development, likely assisting in regulating transcriptional activation and suppression. This category includes the genes *ceh-20*, *ceh-49*, *ceh-51*, *grh-1*, and *lin-1*. *ceh-51* has actually proven important for muscle specification in the MS lineage {Broitman-Maduro, 2009 #63}. Further study on the exact role of the other genes in body wall muscle development may prove worthwhile and insightful. Another

category, including *tbp-1*, *sex-1*, and *sdc-1*, is involved in transcriptional machinery and may reflect the necessity of transcription functioning properly in a crippled network. A third category consists of genes that are known to function in other developmental networks and had previously been believed to play no role in muscle development. These genes, including *cnd-1* and the previously mentioned multifunctional *sex-1* and *sdc-1*, may have unknown roles in muscle differentiation. Alternatively, they may be activated under the extenuating circumstances of a major mutation either to compensate for the mutation or by accident. These genes may prove to be ideal targets for further study on the nature of cross-network interactions and network fate specification. Understanding how these genes can rescue muscle differentiation will go a long way toward understanding how all the components of a network are selected and activated. The final category of genes consists of transcription factors whose function is unknown and their role in muscle development remains to be discovered. These genes will be ideal for small, focused studies on differentiation and single gene dependencies.

As we have expanded our repertoire of regulatory elements and transcription factors, the next logical step was to determine where a specific factor would interact with the genome and to what extent that interaction would be functional. We have demonstrated, through anti-HLH-1 ChIP-seq, where the transcription factor HLH-1 binds and what impact it may have based on the nature and function of nearby genes. Binding is far more prevalent than would be suggested by its functionality, but equally odd is a dearth of binding around certain muscle genes. This reflects that the factor HLH-1 binds almost indiscriminately across the genome but is not required for a significant portion of muscle expression. Much of its regulation may be shared with other factors or indirect.

Since there is a significant level of binding near genes that have nothing to do with muscle development or maintenance, the question arises as to why the factor binds. Since we were able to recapitulate the previously identified E-box binding motif from the various *in vivo* binding sites, we know that HLH-1 preferentially binds to a hexamer. Since such short sequences arise by chance very frequently in a genome, it is possible that most of the binding sites are simply chance binding. Further studies investigating where there is a correlation between the functionality of the binding site and its level of conservation should prove useful and informative. We can predict that non-functional binding sites will not be conserved between species while the developmentally necessary sites will be very well conserved. Further ChIP-seq analyses with other transcription factors, such as the other major muscle factor UNC-120 should be equally informative. To further understand the differences between functional and non-functional binding a combination of anti-RNA Polymerase II ChIP-seq and ChIA-pet should provide the distinction between transcriptional functional and non-functional binding. This may also answer the question of whether HLH-1 serves as a recruitment factor by attracting transcriptional machinery or an initiation/elongation factor that activates transcription once the machinery is assembled.

With a genome-wide understanding of one transcription factor's binding, we sought to understand the functional role that binding plays within the network. This plays into our larger goal of understanding how the different components interact to form a network: what does each component do and what does it not do? With RNA-seq we captured and quantified the transcriptome in the developing embryo. We then generated a muscle-specific gene set by using RNAi knock-downs to increase the proportion of

muscle specified in some animals. To investigate the role that *hll-1* plays within the body wall muscle differentiation network, we compared *hll-1* mutants with wild type animals. Surprisingly, only a subset of muscle genes actually depends on *hll-1* for expression and even fewer completely require it. Much of the HLH-1 binding does correlate with HLH-1 dependent expression. Over half of the genes that have reduced expression in the mutant also have an HLH-1 binding site nearby and are likely directly activated by HLH-1. Overall the role of *hll-1* within the network is tempered by a shared responsibility of activating target genes. The other players include known myogenic factors like *unc-120* and a compensation network. In the mutant *hll-8* is actually upregulated along with much of the non-striated muscle differentiation network. Based on our expression data, the synthetic lethal screen, and HLH-1 binding sites we propose that in the absence of *hll-1* the inhibitory factor *sup-12* is no longer expressed, leading to a transcriptional cascade that results in *hll-8* expression. This, in turn, leads to the activation of the non-striated muscle differentiation network. It is unknown whether this is motivated by an effort to repair the network or is an accidental shadow of the specification process. It is expected that such compensation would only occur in the absence of *hll-1* and that *unc-120* mutation would require a different response. This turns out to be true, as *hll-8* is not upregulated in *unc-120(st364)*. Further studies specifically on *unc-120* mutants will exponentially increase the understanding of the network. The different compensatory mechanism and the potentially different regulatory coverage should be very productive and informative.

In the entirety of this research, we have significantly improved the prediction of cis-regulatory elements, identified additional myogenic factors, determined a

transcription factor's binding profile, and parsed the HLH-1 dependent portion of the body wall muscle differentiation network. This research should have a lasting impact on our understanding of both muscle regulation and gene regulatory networks as a whole.

Appendices

Development 136, 2735-2746 (2009) doi:10.1242/dev.038307

The NK-2 class homeodomain factor CEH-51 and the T-box factor TBX-35 have overlapping function in *C. elegans* mesoderm development

Gina Broitman-Maduro¹, Melissa Owraghi^{1,2,*}, Wendy W. K. Hung^{1,2,*}, Steven Kuntz³, Paul W. Sternberg³ and Morris F. Maduro^{1,†}

The *C. elegans* MS blastomere, born at the 7-cell stage of embryogenesis, generates primarily mesodermal cell types, including pharynx cells, body muscles and coelomocytes. A presumptive null mutation in the T-box factor gene *tbx-35*, a target of the MED-1 and MED-2 divergent GATA factors, was previously found to result in a profound decrease in the production of MS-derived tissues, although the *tbx-35(-)* embryonic arrest phenotype was variable. We report here that the NK-2 class homeobox gene *ceh-51* is a direct target of TBX-35 and at least one other factor, and that CEH-51 and TBX-35 share functions. Embryos homozygous for a *ceh-51* null mutation arrest as larvae with pharynx and muscle defects, although these tissues appear to be specified correctly. Loss of *tbx-35* and *ceh-51* together results in a synergistic phenotype resembling loss of *med-1* and *med-2*. Overexpression of *ceh-51* causes embryonic arrest and generation of ectopic body muscle and coelomocytes. Our data show that TBX-35 and CEH-51 have overlapping function in MS lineage development. As T-box regulators and NK-2 homeodomain factors are both important for heart development in *Drosophila* and vertebrates, our results suggest that these regulators function in a similar manner in *C. elegans* to specify a major precursor of mesoderm.

KEY WORDS: Mesoderm, *C. elegans*, *tbx-35*, *ceh-51*, Tissue specification

INTRODUCTION

During metazoan development, embryonic cells must select from among multiple possible fates, and, ultimately, their descendants will produce gene products typical of a differentiated tissue. In the nematode *C. elegans*, early embryonic cells acquire transient, distinct identities after the zygote undergoes a series of asymmetrical cleavages. These form the six so-called 'founder cells', each of which undergoes a stereotyped pattern of cell divisions to give rise to a nearly invariant set of descendants (Fig. 1A) (Sulston et al., 1983). The emergent paradigm of blastomere/lineage specification is that maternal factors first specify blastomere identity by zygotic activation of blastomere-specific factors, which ultimately leads to activation of tissue-specific gene networks (Labouesse and Mango, 1999; Lei et al., 2009; Maduro, 2009). Blastomere-specific factors are transiently expressed and act for a short time in development, whereas tissue-specific factors tend to maintain their expression throughout the lifespan. An understanding of how lineage-specific activation of tissue factors is achieved will close the gap between studies of blastomere fate and studies of tissue identity, generating a comprehensive gene network that describes development.

The 7-cell stage MS blastomere generates many mesodermal cell types, including cells of the pharynx and body musculature (Fig. 1A,C). The gene cascade that specifies MS has been studied for almost two decades (Fig. 1B). Initial specification of MS

requires maternal activity of the bZIP/homeodomain factor SKN-1 (Bowerman et al., 1993; Bowerman et al., 1992). Loss of *skn-1* leads to a lack of MS-derived tissues and a somewhat less penetrant loss of endoderm from E, the sister cell of MS (Bowerman et al., 1992). *skn-1* mutants also lack the AB-derived portion of the pharynx owing to failure of a Notch/GLP-1-mediated induction from MS to the AB lineage (Priess et al., 1987; Shelton and Bowerman, 1996). In *skn-1* mutants, mis-specified MS and E cells adopt the fate of the mesectodermal precursor C (Bowerman et al., 1992). In EMS (the mother of MS and E), SKN-1 activates the zygotic *med-1 med-2* (*med-1,2*) divergent GATA factor gene pair (Coroian et al., 2005; Maduro et al., 2001). Loss of *med-1,2* has a similar effect on MS specification as loss of *skn-1*, but a much weaker effect on E specification owing to parallel contributions to endoderm from SKN-1 and other factors (Goszczynski and McGhee, 2005; Maduro et al., 2005a; Maduro et al., 2001). In MS, MED-1,2 activate the T-box factor gene *tbx-35* (Broitman-Maduro et al., 2006). Loss of *tbx-35* has variable effects on MS lineage development and morphogenesis, although the most severely affected mutants resemble *skn-1* or *med-1,2* embryos and lack most tissues made by MS (Broitman-Maduro et al., 2006).

The regulatory cascade initiated by SKN-1 works combinatorially with other factors that restrict MS fate to the appropriate blastomere. Within the EMS lineage, SKN-1 and its target genes collaborate with the Wnt/ β -catenin asymmetry pathway to distinguish MS and E identity (Maduro et al., 2002; Rocheleau et al., 1997; Shetty et al., 2005; Thorpe et al., 1997). EMS receives an induction from its posterior sister P₂ that ultimately results in differential nucleocytoplasmic localization of the nuclear effector TCF/POP-1 within MS and E, referred to as POP-1 asymmetry (Goldstein, 1992; Lin et al., 1998; Lo et al., 2004; Maduro et al., 2005a; Rocheleau et al., 1999). Within the E cell, reduced nuclear POP-1 permits POP-1 to function as an

¹Department of Biology, University of California, Riverside, CA 92521, USA.

²Graduate Program in Cell, Molecular and Developmental Biology, University of California, Riverside, CA 92521, USA. ³Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA.

*These authors contributed equally to this work

†Author for correspondence (e-mail: mmaduro@ucr.edu)

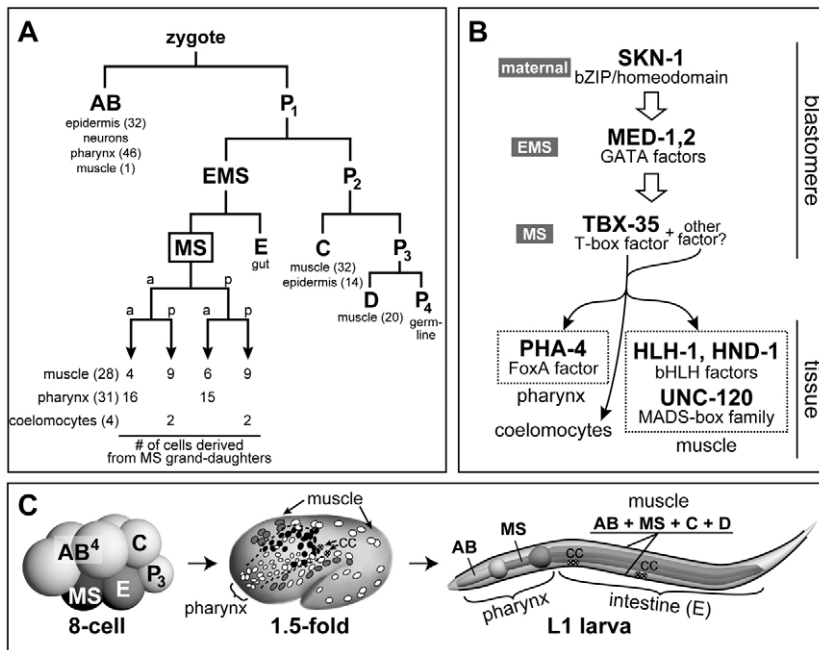


Fig. 1. Developmental context of the MS lineage and its gene regulatory network.

(A) Partial cell lineage showing the production of major tissue types (number of cells in brackets) from early blastomeres (Sulston et al., 1983). The MS lineage is expanded to show the origin of pharynx, muscle and coelomocytes. (B) Gene regulatory network for MS specification [modified with permission from Maduro (Maduro, 2009)]. (C) Embryo stages. Blastomeres are indicated on the 8-cell stage embryo. In the 1.5-fold embryo, all pharynx nuclei, and body muscle nuclei of the left half of the embryo, are shown. Darker-shaded nuclei are those derived from MS. The left-side embryonic coelomocytes (cc) are shown as circles with an X. For the L1 larva, tissues are indicated along with their blastomere of origin. A *C. elegans* embryo is ~50 μ m long. Here and in subsequent figures, anterior is to the left and dorsal is up.

endoderm activator through association with the divergent β -catenin SYS-1 (Huang et al., 2007; Phillips et al., 2007). Blockage of the induction, or of the components that act upstream of POP-1, results in EMS dividing to produce two MS-like cells (Goldstein, 1992; Rocheleau et al., 1997; Rocheleau et al., 1999; Thorpe et al., 1997). Outside of the EMS lineage, multiple factors block inappropriate expression of SKN-1 or prevent its timely degradation, either of which can otherwise lead to ectopic mis-specification of MS or E fates (Lin, 2003; Mello et al., 1992; Page et al., 2007; Shirayama et al., 2006).

The organ-identity factors that specify the two major tissues made by MS, pharynx and muscle, have been well characterized. Pharynx is specified by FoxA/PHA-4 (Horner et al., 1998; Kalb et al., 1998), which is at the top of a network of at least several hundred genes (Gaudet and Mango, 2002) that includes the pharynx muscle-specific gene *ceh-22* (Okkema and Fire, 1994). Body muscle is specified by the activity of three regulators, MyoD/HLH-1, HAND/HND-1 and SRF/UNC-120 (Fukushige et al., 2006). All three genes have overlapping function, as each can specify muscle fate when overexpressed and muscle specification is blocked only when the activity of all three has been compromised (Fukushige et al., 2006). Approximately 1300 genes are known to be enriched for expression in muscle (Fox et al., 2007), suggesting that HLH-1, HND-1 and UNC-120 are at the top of a complex tissue-specific muscle gene network.

In the present study we identify the NK-2 homeobox gene *ceh-51* as a direct target of TBX-35, and present evidence that CEH-51 and TBX-35 have distinct and shared functions. Whereas loss of *ceh-51* function causes subtle muscle and pharynx defects and larval lethality, simultaneous loss of *ceh-51* and *tbx-35* results in a highly penetrant loss of MS-derived tissues and an embryonic arrest phenotype that is strikingly similar to that of *med-1,2(-)* embryos, thus explaining the weaker phenotype of single *tbx-35* mutants. Our results add an important regulator, CEH-51, to the MS gene regulatory network, and suggest that combinatorial control of mesoderm through T-box and NK-2 factors has been evolutionarily conserved.

MATERIALS AND METHODS

Strains used

C. elegans animals were cultured on *E. coli* OP50 using standard methods (Sulston and Hodgkin, 1988). The wild-type strain was N2. Mutations: LG X: *hnd-1(q740)*, *med-1(ok804)*. LG I: *unc-120(st364)*. LG II: *tbx-35(tm1789)*, *hlh-1(cc561ts)*. LG III: *unc-119(ed4)*, *med-2(cx9744)*. LG IV: *skn-1(zu67)*. LG V: *ceh-51(tm2123)*. Rearrangement: *nT1 [unc-?(n754) let-?(IV;V)]*. Transgenes: *gvlS401 V [unc-120::GFP]*, *gvlS402 I [unc-120::GFP]*, *culs1 V [ceh-22::GFP]*, *ccls7963 V [hlh-1::GFP]*, *qls55 [hnd-1::GFP]*, *irIs57 III [hs-ceh-51]*, *irIs70 [hs-ceh-51]*, *cdIs41 II [cup-4::GFP]*, *cdIs42 I [cup-4::GFP]*, *ruIs37 III [myo-2::GFP]*, *pxIs[pha-4::GFP] IV*, *irIs39 III [ceh-51::GFP]*, *irIs41 [ceh-51::GFP]*, *irIs42 X [hs-tbx-35]*, *irIs58 [hs-ceh-51]*, *irIs89 [ceh-51(+)]*, *qtlS9 [nhr-25::YFP]*. We have previously observed a lack of strict additivity and variability in the number of cells expressing tissue-specific reporters (Lin et al., 2009). We attribute this primarily to expression mosaicism between animals and the difficulty of resolving adjacent cells.

Identification of *ceh-51*

Y80D3A.3 (previously *dlx-1*) was named *ceh-51* in consultation with Thomas Burglin and Jonathan Hodgkin (Karolinska Institute, Stockholm, Sweden and University of Oxford, Oxford, UK). *ceh-51* resides within intron 12 of Y80D3A.2/*emb-4* (WormBase, WS200 release). Four ESTs support a single transcript with one intron for *ceh-51* that does not overlap *emb-4* exonic sequences (Kohara, 2001). As RNAi targeted to introns does not affect mature transcripts (Fire et al., 1998), it is unlikely that RNAi targeted to *ceh-51* would affect transcripts of *emb-4*. Indeed, RNAi of *emb-4* results in embryonic lethality (Katic and Greenwald, 2006), not larval arrest (see text).

Construction of *ceh-51(tm2123)* strains

We injected a heterozygous *ceh-51(tm2123)* strain (a gift from Shohei Mitani, National Bioresource Project, Japan) with overlapping genomic PCR products spanning the *ceh-51* locus (but lacking any exonic *emb-4* sequences; primer sites are shown in Fig. 2) and an *unc-119::CFP* reporter (pMM809) to produce MS1206, a line that segregated arrested larvae and *unc-119::CFP(+)* viables. We confirmed the correct splicing of *emb-4* in the *tm2123* strain by RT-PCR. After backcrossing, the array was replaced with another carrying *ceh-51(+)*, *unc-119::mCherry* (pMM824) and *myo-2::mCherry* (pCFJ90) for the muscle phenotype synergy experiments. PCR confirmed homozygosity of the *tm2123* deletion in this strain. A

spontaneous integrant of a *ceh-51(+)* array, *irIs89*, showed that 96% ($n=253$) of *ceh-51(tm2123); irIs89* embryos were rescued to full viability. A *tbx-35(tm1789); ceh-51(tm2123)* double mutant strain was made by crossing *tbx-35; Ex[tbx-35(+), unc-119::YFP]* males to *ceh-51; Ex[ceh-51(+), unc-119::CFP]* hermaphrodites, and identifying YFP/CFP-expressing F₂ animals that gave arrested embryos/larvae and in which all viable animals expressed both YFP and CFP. The two arrays in MS1275 were replaced by a single array marked with *unc-119::mCherry* (pMM824) or *sur-5::dsRed* (pAS152).

Cloning and transgenics

To construct *ceh-51::GFP* (pGB196), a PCR product containing 788 bp upstream of the *ceh-51* start codon and 204 bp of the coding region was cloned into the *SphI-BamHI* sites in pPD95.67. A smaller reporter, with 187 bp of upstream DNA and 5 bp of coding region, was cloned similarly (pWH270). TBX-35 sites were mutated into restriction sites by PCR in pWH270. A translational fusion was constructed by combining 358 bp of *ceh-51* promoter, a GFP coding region from pPD95.67 and the genomic region of *ceh-51* containing the exons, intron and 3'UTR. A heat-shock *ceh-51* construct was created by cloning the coding region, intron and 468 bp of the 3'UTR into pPD49.78. Further PCR and cloning details are available on request. Transgenics and integrants were made as described (Maduro et al., 2001).

RNAi experiments

For feeding-induced RNAi, L4 animals were fed for 36 hours on *E. coli* HT115 from the OpenBioSystems RNAi Library or transformed with clones made in pPD129.36. Adults were transferred to fresh plates for egg laying for 4-6 hours at 25°C. Embryos were allowed to develop for 12-24 hours prior to scoring. For dsRNA synthesis, PCR products carrying the T7 RNA polymerase recognition sequence at each end were amplified from N2 DNA, cDNA clones or the Ahringer Lab RNAi Library (Kamath and Ahringer, 2003). dsRNA was synthesized using the Ambion MEGAscript T7 Kit and microinjected into late L4 worms or young adults as described (Ahringer, 2006). Injected animals were allowed to recover for 3-24 hours and transferred to fresh plates for egg laying.

In situ hybridization

Embryos were stained as described (Coroian et al., 2005). For *pal-1* staining of *med-1,2* and *ceh-51; tbx-35* embryos, a mixture of rescued and non-rescued embryos were stained, and the number of mutants was estimated from the array transmission frequency.

Phalloidin staining

Embryos or larvae were freeze-cracked on dry ice or frozen in liquid nitrogen, fixed in 4% formaldehyde and stained with Alexa Fluor 488-conjugated phalloidin (Molecular Probes, Eugene, OR, USA) as described (Shaham, 2006).

Laser ablation, microscopy and imaging

Laser ablations were performed as described (Lin et al., 2009). Animals were imaged on a Zeiss Axioplan using a Hamamatsu ORCA II digital camera, or on an Olympus BX-61 with a Canon 350D camera. For phalloidin-stained larvae, a Zeiss LSM510 confocal microscope was used (Microscopy and Imaging Core Facility, UC Riverside). Adobe Photoshop 7 and ImageJ v1.37 were used to adjust image brightness and generate overlays.

Heat-shock experiments

Embryos were heat shocked as a group for 30-45 minutes at 33°C while they were contained within hermaphrodite mothers, representing a developmental time interval of 0-3 hours. After heat shock, hermaphrodites were allowed to lay eggs for 3-4 hours. Embryos were allowed to develop for a further 6-12 hours before scoring. For in situ hybridizations after heat shock, mothers were left overnight at 15°C on plates without food.

Expression and purification of recombinant TBX-35

A cDNA fragment encoding amino acids 120-325, corresponding to the predicted TBX-35 DNA-binding domain, was cloned into the GST vector pGEX-4T-1 (GE Healthcare) to generate pWH173. This was transformed into *E. coli* Rosetta2 cells (Novagen), grown at 37°C to an OD of 0.3, and

protein production was induced with 0.3 mM IPTG overnight at 25°C. Cells were resuspended in BugBuster HT (Novagen) with one tablet of Complete, Mini, EDTA-free Protease Inhibitor (Roche). Glutathione beads, swelled in phosphate-buffered saline (PBS), were added to the lysate for 1 hour. After three washes with PBS, the protein was eluted in 50 mM Tris-HCl (pH 7.5), 5 mM reduced glutathione, 80 mM NaCl, 0.03% Triton X-100, and desalted using a P6 column (BioRad). The protein was stored at -20°C in 50% glycerol with 1 mM DTT and 10 mM Tris (pH 7.5).

Gel shift and DNase I footprinting

EMSA probes were gel-purified PCR products generated with a ³²P end-labeled primer and an unlabeled primer. The probes contained DNA corresponding to -187 bp to +5 bp relative to the *ceh-51* ATG. Probes carrying mutated sites were amplified from the corresponding GFP reporters. Gel shift and DNase I footprinting were performed as previously described for MED-1 (Broitman-Maduro et al., 2005), except that 10 μM GST and 10, 25 and 50 μM GST::TBX-35(DBD) were used, 6% acrylamide gels were run, and complexes were treated with 0.5 units of DNase I (Epicentre) for 40 seconds prior to organic extraction. For competition arrays, complementary oligonucleotides were annealed at 95°C for 5 minutes, cooled for 15 minutes and added to reactions at a 50-fold excess.

RESULTS

Identification of CEH-51, a putative NK-2 class homeodomain transcription factor

Loss of *med-1,2* leads to a highly expressive loss of MS-derived tissues, whereas loss of *tbx-35* has a less expressive MS phenotype, especially at lower temperatures (Broitman-Maduro et al., 2006; Maduro et al., 2001) (this work), suggesting that an additional factor contributes to MS specification downstream of MED-1,2 (Broitman-Maduro et al., 2006). From embryonic transcriptome analyses (Baugh et al., 2005; Baugh et al., 2003), we identified Y80D3A.3 as a candidate early MS lineage gene. Transcripts were reported to accumulate when the MS lineage is undergoing its first divisions, and were reduced in *mex-3(zu155); skn-1(RNAi)* embryos, which do not correctly specify MS. In parallel, we identified Y80D3A.3 in an RNAi screen for enhancement of *hlh-1(cc561ts)* muscle defects (S.K. and P.W.S., unpublished results).

The Y80D3A.3 gene encodes a putative homeodomain transcription factor, CEH-51 (Fig. 2). Of the 89 homeodomain proteins encoded by the *C. elegans* genome (Okkema and Krause, 2005), CEH-51 is most closely related to CEH-7 (Kagoshima et al., 1999), CEH-24 (Harfe and Fire, 1998) and TAB-1 (CEH-29) [L. Carnell and M. Chalfie, unpublished data cited in Syntichaki and Tavernarakis (Syntichaki and Tavernarakis, 2004)], sharing 41-48% identity (57-58% similarity) within the homeodomain (Fig. 2B,C). The CEH-51 homeodomain is most closely related to those of NK-2 subfamily proteins, with which it shares 39-43% identity (59-62% similarity), although CEH-51 lacks the conserved tyrosine at position 54 of the homeodomain (asterisk in Fig. 2C) that is typical of NK-2 proteins (Harvey, 1996). The *C. elegans* pharynx muscle NK-2 factor CEH-22 is more closely related to other NK-2 family members, as it contains the conserved tyrosine and shares 85% identity (90% similarity) with *Drosophila* Vnd/NK-2 across the homeodomain. CEH-51 contains multiple serine residues in its N-terminus (16/50 residues), a feature noted for the N-termini of CEH-24 (Harfe and Fire, 1998) and the endoderm-specifying END-1,3 GATA factors (Maduro et al., 2005b).

ceh-51 is expressed in the early MS lineage downstream of TBX-35

We confirmed that *ceh-51* transcripts accumulate in the MS daughters and persist into the MS granddaughters, as observed in 91% ($n=70$) of embryos at the MS² to MS⁴ stage (Fig. 3A,B).

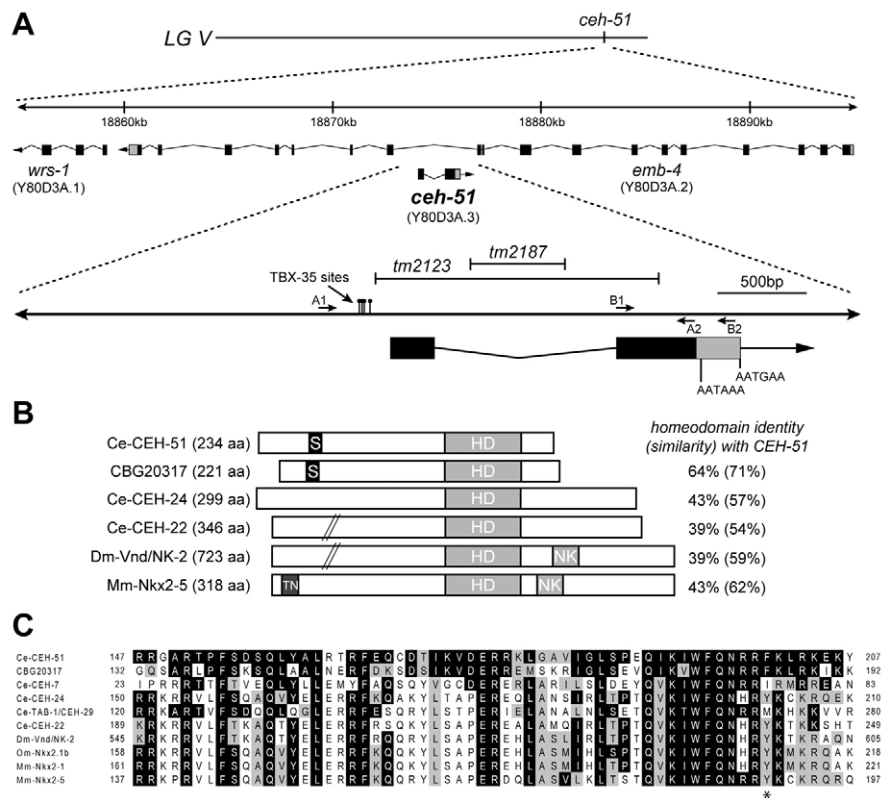


Fig. 2. Structure of *ceh-51* and its gene product. (A) Location of *ceh-51* and *emb-4* exons on LG V. The locations of the mutant alleles *tm2123* and *tm2187* and the primer pairs A1/A2 and B1/B2 (used to generate overlapping PCR products for rescue of *tm2123*) are shown. A 3'UTR of ~260 bases is predicted by EST yk51g7. Polyadenylation motifs of AATAAA and AATGAA (Hajarnavis et al., 2004) are found 40 bp and 260 bp, respectively, downstream of the stop codon. *tm2123* is a 1610 bp deletion that includes the coding portion of exon 1 and part of exon 2, including the first six amino acids of the predicted homeodomain, and carries an additional 14 bp insertion. The remainder of the *ceh-51* coding region in *tm2123* lacks any in-frame ATG codons, suggesting that *tm2123* is null. *tm2187* is an intronic 540 bp deletion and was not studied. (B) Comparison of CEH-51 and other NK-2 factors. Like all *C. elegans* NK-2 factors, CEH-51 lacks the Tinman (TN) and NK-2-specific (NK) domains that are found in many other NK-2 factors (Harvey, 1996). Regions where at least 7/10 contiguous residues are serine are indicated by S. HD, homeodomain. (C) Homeodomain alignments. Identities with *C. elegans* CEH-51 are indicated by black boxes and similarities by gray boxes. A tyrosine residue found in NK-2 family members is indicated with an asterisk (Harvey, 1996). Accession numbers: *C. elegans* (Ce) CEH-51, CAB60440; CEH-7, AAC36745; CEH-24, AAB81844; TAB-1 (CEH-29), AAA98021; CEH-22, NP_001076744; *C. briggsae* CBG20317, CAP37360; *Oncorhynchus mykiss* (Om) Nkx2.1b, BAD93686; *Mus musculus* (Mm) Nkx2.1, NP_033411; Nkx2.5, NP_032726; *Drosophila melanogaster* (Dm) Vnd, P22808.

Similar expression was seen with a *ceh-51::GFP* transcriptional reporter carrying 788 bp of genomic DNA upstream of the predicted ATG (Fig. 3E), a GFP::CEH-51 translational fusion with 358 bp of upstream region (Fig. 3F), and from expression reported by others (Hunt-Newbury et al., 2007; Kohara, 2001; Reece-Hoyes et al., 2007). As anticipated by the mis-specification of MS in *skn-1* and *med-1,2* mutant embryos (Bowerman et al., 1992; Maduro et al., 2001), expression of *ceh-51::GFP* was not observed in these backgrounds (Fig. 3G; data not shown). Conversely, ectopic *ceh-51::GFP* was observed in *mex-1* and *pie-1* RNAi backgrounds (Fig. 3I,J), in which additional MS-like cells are made from the AB and C lineages, respectively (Mello et al., 1992). We have previously found that *tbx-35* is still expressed in MS in a *pop-1(RNAi)* background (Broitman-Maduro et al., 2006), even though in this background MS adopts an E-like fate (Lin et al., 1995). Unexpectedly, most *pop-1(RNAi)* embryos expressed *ceh-51* in both the MS and E lineages (Fig. 3D,H).

The expression pattern of *ceh-51* suggests that it is a direct target of TBX-35. Overexpression of TBX-35 was sufficient to cause ectopic *ceh-51* activation (Fig. 3C), whereas weaker expression still occurred in approximately half of *tbx-35(tm1789)* mutants (Fig. 3K),

demonstrating that TBX-35 is sufficient but not necessary for *ceh-51* activation. In a *tbx-35(tm1789); pop-1(RNAi)* background, expression of *ceh-51::GFP* became undetectable (Fig. 3L), suggesting that activation of *ceh-51* in a *tbx-35* mutant background is POP-1-dependent.

To test for direct interaction of TBX-35 with *ceh-51*, we purified recombinant GST::TBX-35 DNA-binding domain (DBD) expressed in *E. coli*, and found that a 187 bp fragment of *ceh-51* could be gel shifted (Fig. 4A, lanes 6-8). We identified four putative TBX-35 binding sites based on similarity to the consensus sequence for the founding T-box factor Brachyury (Kispert and Herrmann, 1993), and confirmed that they were protected in a DNase I footprinting assay (Fig. 4B). These regions define a consensus of RTSKCACCCYNNYY (Fig. 4C), which matches 7/8 sites of the Brachyury half-site TCACACCT (matches underlined) (Kispert and Herrmann, 1993). Hence, it is likely that TBX-35 binds DNA as a monomer, similar to mouse Tbx20 and Tbx5 (Ghosh et al., 2001; Macindoe et al., 2009; Stennard et al., 2003). A competitor oligonucleotide containing two of the candidate sites competed the shifts, whereas a competitor with both sites mutated did not (Fig. 4A, lanes 9-11), and all four sites appear to be important for TBX-35

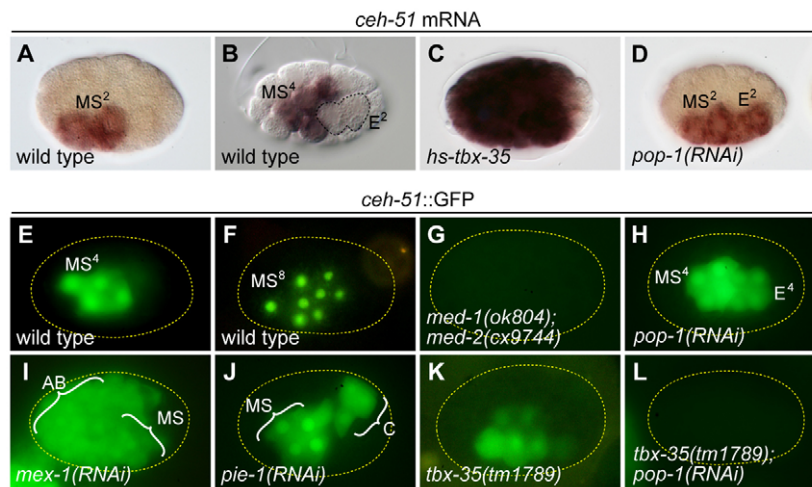


Fig. 3. Expression of *ceh-51*. (A,B) *ceh-51* transcripts occur in (A) the MS daughter cells (MS²) and (B) in the MS granddaughters (MS⁴), as detected by in situ hybridization. The E daughters are outlined. Ninety-one percent ($n=101$) of embryos at this stage showed expression in MS² or MS⁴ (nine embryos did not stain). (C) Ectopic expression of *ceh-51* following heat shock of *hs-tbx-35* embryos. (D) Eighty-six percent ($n=44$) of *pop-1(RNAi)* embryos showed *ceh-51* mRNA in both the MS and E daughters. Two embryos showed normal expression and four embryos did not stain. (E) Embryos transgenic for a *ceh-51::GFP* transcriptional reporter with 788 bp of upstream sequence show expression at MS⁴ that persists in later MS descendants. (F) A translational *ceh-51::GFP::CEH-51* fusion shows strong nuclear accumulation at MS⁸. (G) *ceh-51::GFP* is undetectable in *med-1(ok804); med-2(cx9744)* embryos ($n=84$). (H) Sixty-six percent ($n=41$) of *pop-1(RNAi)* embryos showed *ceh-51::GFP* in both the MS and E lineages (the remainder were similar to wild type). (I,J) *mex-1(RNAi)* (I) and *pie-1(RNAi)* (J) embryos displayed ectopic *ceh-51::GFP* in AB and C descendants. (K) In *tbx-35(tm1789)* embryos, the onset of *ceh-51::GFP* expression was undetectable (52%, $n=89$) or delayed until past the MS⁸ stage (48%) and at lower levels. The exposure in this image was 10-fold longer than that shown in E. (L) *ceh-51::GFP* was not detected in *tbx-35(tm1789); pop-1(RNAi)* embryos ($n=49$).

binding (Fig. 4D). In vivo, a minimal *ceh-51::GFP* reporter carrying the four sites was expressed in the early MS lineage, its expression was abolished in a *tbx-35(tm1789)* background, and mutation of the sites resulted in a loss of expression (Fig. 4E). We conclude that TBX-35 directly activates *ceh-51*.

Overexpressed CEH-51 is sufficient to promote aspects of MS specification

We next assessed the ability of CEH-51 to specify the development of MS-derived cell types using a heat-shock (hs) *ceh-51* transgene. Ninety-one percent ($n=245$) of heat shocked pregastrulation *hs-ceh-51* embryos underwent arrest, whereas heat shock of wild types resulted in only 22% ($n=243$) embryonic arrest. We examined pharynx muscles with *ceh-22::GFP* (Okkema and Fire, 1994), using a *skn-1(RNAi)* background to eliminate MS-derived tissues and AB-derived pharynx (Bowerman et al., 1992). Among *skn-1(RNAi); hs-ceh-51* embryos, we observed only a small number of *ceh-22::GFP*-positive cells following heat shock (Fig. 5F), and were unable to detect significant expression of the pharynx identity gene *pha-4* (Horner et al., 1998) or the pharyngeal myosin gene *myo-2* (Miller et al., 1986) (Fig. 5G; data not shown), suggesting that CEH-51 by itself has, at most, a weak ability to specify pharynx.

Next, we examined production of muscle in a *skn-1(RNAi); pal-1(RNAi)* background, which blocks specification of nearly all body muscles (Hunter and Kenyon, 1996). In such embryos, *hs-ceh-51* was sufficient to promote widespread muscle specification as scored by *unc-120::GFP* (Fukushige et al., 2006) and expression of the body muscle gene *myo-3* (Miller et al., 1986) (Fig. 5H,I). Hence, CEH-51 is sufficient to specify muscle cell fate.

We then examined production of the four embryonically derived coelomocytes, which arise fairly late in the MS lineage (Sulston et al., 1983), using *cup-4::GFP* (Patton et al., 2005). *hs-ceh-51* was

sufficient to cause specification of coelomocytes in a *skn-1(RNAi)* background, which by itself eliminates them (Table 1; Fig. 5E,F). We conclude that CEH-51 is sufficient to specify muscle and coelomocyte precursors. No attempt was made to optimize the time interval for CEH-51 responsiveness, although under the same conditions, overexpressed *tbx-35* was able to cause specification of pharynx, muscle and coelomocytes (Broitman-Maduro et al., 2006) (data not shown).

Loss of *ceh-51* function results in defects in MS-derived tissues

To evaluate the requirement for *ceh-51* in MS specification, we examined *ceh-51(RNAi)* and *ceh-51(tm2123)* animals. Gonadal injection of *ceh-51* dsRNA resulted in 47% ($n=70$) of progeny arresting as uncoordinated L1 larvae, whereas the remainder appeared normal (50%) or arrested as early embryos (3%). The putative null mutant, *tm2123* (Fig. 2A), resulted in a fully penetrant recessive zygotic L1 arrest. This lethality could be rescued by a *ceh-51(+)* transgene (see Materials and methods).

We examined *ceh-51* mutants for pharynx defects. *ceh-51(tm2123)* mutants had a poorly defined metacarpus and an incompletely developed grinder (Fig. 6A,D), and expression of the pharynx muscle reporter *ceh-22::GFP* (Okkema and Fire, 1994) was observed both inside and outside of the pharynx basement membrane, suggesting defective pharynx integrity (Fig. 6B,E). We also observed detachment of the pharynx from the buccal cavity in 64% ($n=56$) of animals. Similar defects were apparent in *ceh-51(RNAi)* arrested larvae (data not shown). *ceh-51(tm2123)* mutants also had defects in the organization of actin filaments as detected by phalloidin staining (Fig. 6C,F). We scored production of all pharynx cells in *ceh-51* mutants using a *pha-4::GFP* reporter (Horner et al., 1998), and found that the number of cells in *ceh-51* mutants

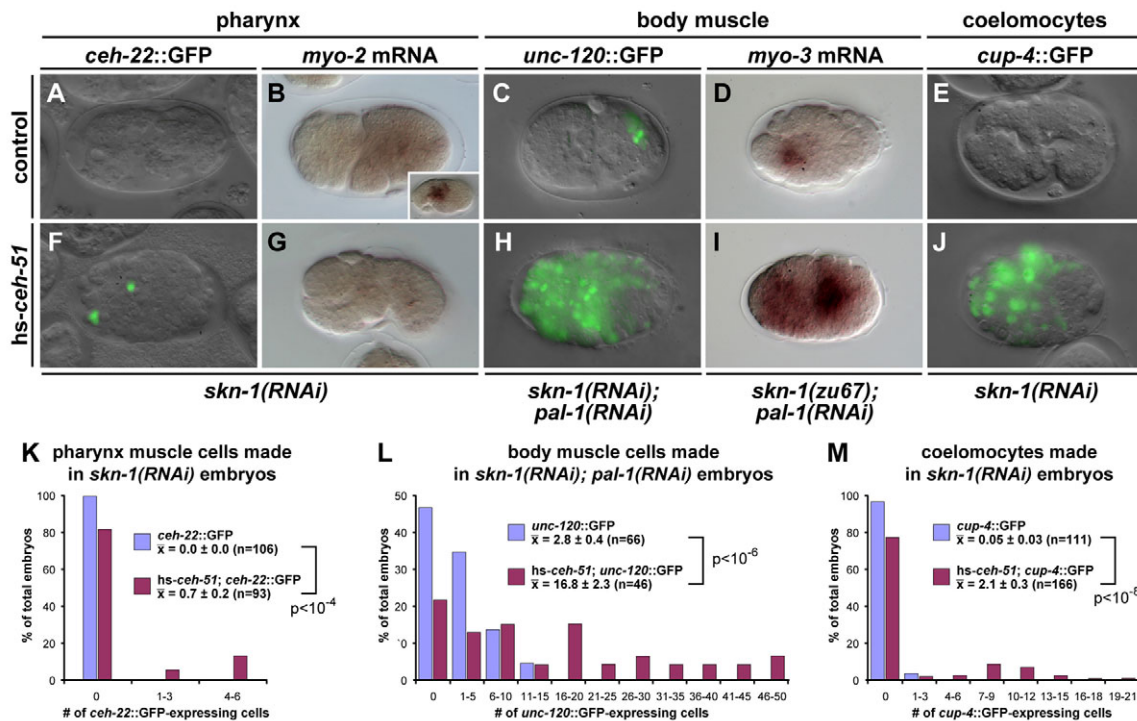


Fig. 5. Overexpression of CEH-51 promotes specification of MS-derived cell types. (A,F) A small number of cells expressing *ceh-22::GFP* are restored to *skn-1(RNAi)* embryos by *hs-ceh-51*. (B,G) Expression of the pharynx muscle gene *myo-2* is largely absent in both *skn-1(RNAi)* and *skn-1(RNAi); hs-ceh-51* embryos. The inset in B shows wild-type expression of *myo-2*. (C,H) Many *hs-ceh-51* embryos display *unc-120::GFP*-expressing cells in a *skn-1(RNAi); pal-1(RNAi)* background, which depletes embryos of nearly all body muscles (Hunter and Kenyon, 1996). (D,I) In a *skn-1(zu67); pal-1(RNAi)* background, heat shock of *ceh-51* causes the generation of many cells expressing the muscle myosin gene *myo-3*. One hundred percent ($n=79$) of heat shocked non-transgenic embryos resembled those shown in D, whereas 53% ($n=53$) of heat shocked transgenics resembled those shown in I. (E,J) *hs-ceh-51* embryos accumulate ectopic coelomocytes. (K-M) Bar charts summarizing the *hs-ceh-51* data.

descendants is compromised in mutants. Taken together, these results show that CEH-51 is required for the normal development of multiple MS tissue types.

TBX-35 and CEH-51 have overlapping function

Loss of *med-1,2* results in an embryonic lethal phenotype in which arrested embryos elongate to between one and two times the length of the eggshell (Maduro et al., 2007; Maduro et al., 2001). By contrast, *tbx-35* null mutants arrest with varying degrees of elongation, ranging from 1-fold to complete elongation and hatching (Broitman-Maduro et al., 2006). These results suggest that TBX-35 works with another factor. Two further observations support this notion. First, whereas *med-1(ok804); med-2(cx9744)* double mutants and *skn-1(RNAi)* embryos made less than 0.2 coelomocytes per embryo (Table 1; Fig. 8L), *tbx-35(tm1789)* embryos raised at 15°C made as many coelomocytes (3.8 ± 0.2 , $n=28$) as wild types (3.7 ± 0.2 , $n=105$, $P > 0.9$). Second, *tbx-35(tm1789)* embryos achieved further elongation overall when raised at 15°C (Fig. 8P). This increased elongation correlated with an increase in production of MS-derived pharynx cells as scored by *pha-4::GFP* in a *glp-1(RNAi)* background, which eliminates AB-derived pharynx (Priess et al., 1987) (Table 1); *tbx-35; glp-1(RNAi)* embryos at 15°C made 6.6 ± 0.5 pharynx cells ($n=23$), whereas at 23°C only 1.1 ± 0.3 cells were made ($n=32$, $P < 10^{-11}$).

We hypothesized that *tbx-35* and *ceh-51* double mutants might show a stronger phenotype than either single mutant, given that *ceh-51* is still activated in *tbx-35(tm1789)* (Fig. 3K). As shown in Fig. 8

and Table 1, *ceh-51(tm2123); tbx-35(tm1789)* double mutants displayed phenotypes that are indistinguishable from *med-1(ok804); med-2(cx9744)* (henceforth abbreviated as *ceh-51; tbx-35* and *med-1,2*). First, *ceh-51; tbx-35* double mutants displayed a strong embryonic arrest that is not temperature sensitive ($P=0.48$ for 15°C versus 20°C) (Fig. 8P) and which is comparable to that of *med-1,2* double mutants at both temperatures ($P=0.36$ and $P=0.43$ for 15°C and 20°C, respectively). Second, development of MS-derived pharynx was eliminated in *ceh-51; tbx-35* (Table 1; Fig. 8E,F), even at 15°C, at which single *ceh-51* and *tbx-35* mutants each displayed a partial grinder (Fig. 6D,G). Using *glp-1(RNAi)* to eliminate AB-derived pharynx, both *med-1,2; glp-1(RNAi)* and *ceh-51; tbx-35; glp-1(RNAi)* embryos made similarly low numbers of pharynx cells (less than two) as scored with *pha-4::GFP* or *ceh-22::GFP* ($P=0.15$ and $P=0.3$) (Table 1). Production of *pal-1*-independent body muscle cells was reduced in *ceh-51; tbx-35; pal-1(RNAi)* embryos to levels comparable to *med-1,2; pal-1(RNAi)* ($P=0.9$) (Table 1; Fig. 8H,I). Lastly, whereas single *ceh-51* and *tbx-35* mutants made reduced numbers of *cup-4::GFP*(+) cells, the double mutants displayed a synergistic reduction similar to that of a *med-1,2* background ($P=0.04$) (Table 1; Fig. 8K,L).

MS adopts a C-like fate in *med-1,2(RNAi)* and *skn-1* mutant embryos (Bowerman et al., 1992; Maduro et al., 2001), but this transformation is weaker in *tbx-35* mutants as zygotic activation of *pal-1* in the MS lineage, a marker of transformation of MS to C (Baugh et al., 2005), was detected in only ~30% of embryos (Broitman-Maduro et al., 2006). We found that 75% ($n=20$) of *med-*

Table 1. MS-dependent tissues produced in wild-type and mutant embryos

Genotype	Pharynx cells [†] (<i>pha-4::GFP</i>)	Pharynx muscles [‡] (<i>ceh-22::GFP</i>)	Muscle cells (<i>h1h-1::GFP</i>)	Coelomocytes (<i>cup-4::GFP</i>)
Wild type	50.0±0.9 (21)	12.8±0.1 (37)	44.7±1.1 (20)	3.7±0.2 (105)
<i>skn-1(RNAi)</i>	4.8±0.4 (20)	0.0±0.0 (165)	nd	0.15±0.04 (124)
<i>pal-1(RNAi)</i>	49.5±0.8 (10)	11.7±0.3 (12)	21.6±0.9 (13)	3.7±0.1 (103)
<i>pop-1(RNAi)</i>	nd	nd	nd	0.0±0.0 (50)
<i>glp-1(RNAi)</i>	23.1±0.6 (15)	5.7±0.2 (38)	nd	nd
<i>tbx-35(tm1789)</i> 15°C	40.6±1.2 (17)	5.9±0.3 (24)	37.3±1.6 (10)	3.8±0.2 (28)
<i>tbx-35(tm1789)</i> 23°C	35.7±0.8 (16)**	5.2±0.2 (46)*	34.8±2.4 (10)	3.3±0.4 (20)
<i>tbx-35(tm1789); glp-1(RNAi)</i> 15°C	6.6±0.5 (23)	2.0±0.4 (26)	nd	nd
<i>tbx-35(tm1789); glp-1(RNAi)</i> 23°C	1.1±0.3 (32)**	1.0±0.2 (39)*	nd	nd
<i>tbx-35(tm1789); pal-1(RNAi)</i> 15°C	38.8±0.7 (15)	5.1±0.3 (14)	8.4±1.0 (17)	2.2±0.2 (47)
<i>tbx-35(tm1789); pal-1(RNAi)</i> 23°C	35.6±1.0 (14)*	4.7±0.3 (17)	5.7±0.5 (40)*	0.6±0.1 (49)**
<i>ceh-51(tm2123)</i>	47.8±0.9 (17)	9.2±0.2 (10)	42.4±1.4 (10)	2.1±0.1 (53)
<i>ceh-51(tm2123); pal-1(RNAi)</i>	nd	nd	19.3±0.5 (11)	2.5±0.1 (84)
<i>med-1(ok804); med-2(cx9744)</i>	31.3±0.6 (26)	4.1±0.2 (32)	31.0±2.7 (10)	0.07±0.03 (34)
<i>ceh-51(tm2123); tbx-35(tm1789)</i>	30.2±0.5 (44)	4.4±0.2 (18)	30.1±1.0 (14)	0.19±0.04 (124)*
<i>med-1(ok804); med-2(cx9744); glp-1(RNAi)</i>	1.4±0.4 (14)	0.3±0.1 (31)	nd	nd
<i>ceh-51(tm2123); tbx-35(tm1789); glp-1(RNAi)</i>	1.9±0.5 (26)	0.5±0.1 (52)	nd	nd
<i>med-1(ok804); med-2(cx9744); pal-1(RNAi)</i>	nd	nd	3.8±0.5 (13)	nd
<i>ceh-51(tm2123); tbx-35(tm1789); pal-1(RNAi)</i>	nd	nd	3.9±0.4 (15)	nd

Strains were grown at 20–23°C unless otherwise indicated. Data are shown as the mean ± s.e.m. *0.01 < P < 0.05, **P < 0.01, Student's t-test, by comparison with the experiment immediately above. nd, not done.

[†]Only pharynx expression of *pha-4::GFP*, anterior to the gut (when present), was scored.

[‡]The anatomy of the pharynx was considered in assigning expression to particular muscle cells.

1,2(–) and 60% (*n*=35) of *ceh-51*; *tbx-35* embryos showed ectopic zygotic *pal-1* mRNA in the early MS lineage (*P*>0.3) (Fig. 8N,O). We examined the fate of MS descendants in *tbx-35*; *ceh-51* double mutants carrying a reporter fusion for *nhr-25*, a C-lineage gene that is expressed in hypodermal precursors and their descendants (Baugh et al., 2005), using a laser to ablate all other cells. Partial embryos resulting from isolated wild-type MS blastomeres failed

to show significant *nhr-25::YFP* (*n*=3), whereas 9/9 MS blastomeres from *tbx-35*; *ceh-51* double mutants, and 5/5 isolated C blastomeres from wild types, generated *nhr-25::YFP* descendants. Hence, *ceh-51*; *tbx-35* embryos show a strong transformation of MS to C, suggesting that CEH-51 and TBX-35 together account for the majority of normal MS lineage development downstream of MED-1,2.

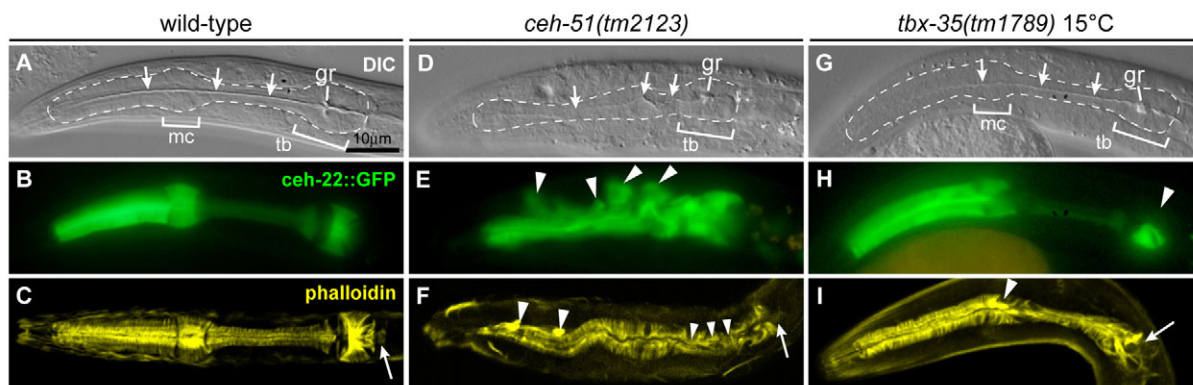


Fig. 6. *ceh-51* mutants and *tbx-35* mutants raised at 15°C arrest as larvae with pharynx structural defects. Pharynxes were visualized by DIC optics (A,D,G), *ceh-22::GFP* expression (B,E,H) (Okkema and Fire, 1994) or phalloidin staining (C,F,I) (Franks et al., 2006). In the DIC panels, the lumen (arrows), grinder (gr), metacarpus (mc) and terminal bulb (tb) are indicated and the pharynx is outlined (dashed line). (A–C) Wild-type pharynx. (D–F) *ceh-51(tm2123)* pharynxes show lumen irregularities and an indistinct metacarpus (D). Protrusions accumulate GFP outside the pharynx, suggesting a defect in pharynx integrity (E). In F, phalloidin staining shows actin filament accumulations (large arrowheads), lumen abnormalities (small arrowheads) and an abnormal terminal bulb (arrow). (G–I) *tbx-35(tm1789)* raised at 15°C has a normal lumen but abnormal grinder (G). *ceh-22::GFP* expression (H) shows absence of expression of *ceh-22::GFP* in part of the posterior pharynx (arrowhead); contralateral expression in this region is likely to be in an MS-derived m7 muscle (Okkema and Fire, 1994; Sulston et al., 1983). In I, phalloidin staining shows some actin accumulations (arrowhead) and an abnormal terminal bulb (arrow).

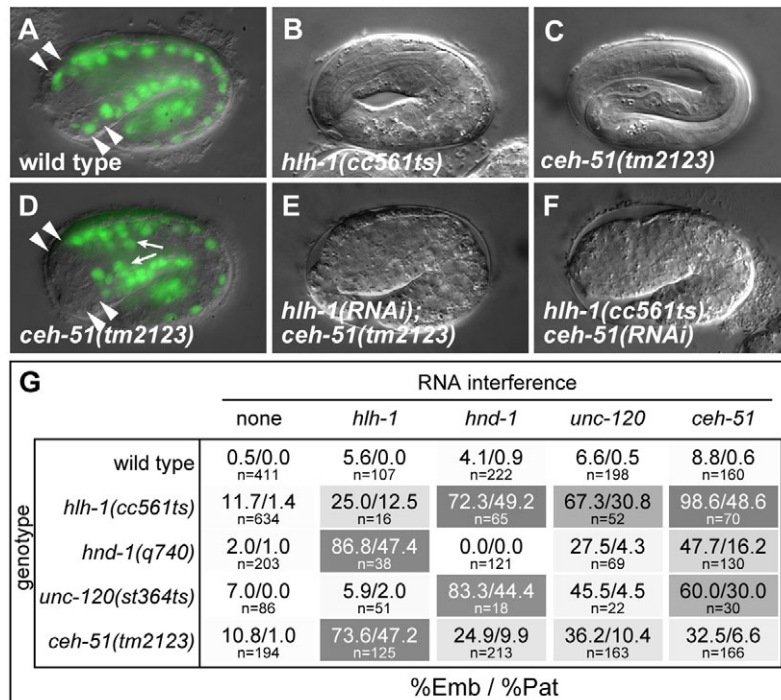


Fig. 7. Muscle defects in *ceh-51(tm2123)*. (A,D) Loss of MS-derived *unc-120::GFP* expression (arrowheads) in *ceh-51* (D) as compared with wild type (A). Additional expression is indicated by small arrows. (B,C,E,F) Loss of *ceh-51* synergizes with partial muscle specification mutants to produce paralyzed, arrested 2-fold (Pat) embryos. Whereas more than 95% of *hlh-1(cc561ts)* mutants grown at 15°C (B), and more than 99% of *ceh-51(tm2123)* embryos (C), elongated to greater than 3-fold, between 47 and 49% of embryos produced by a combination of mutation of *ceh-51* with RNAi of *hlh-1* (E), or vice versa (F), produced a synthetic Pat phenotype. (G) Summary of synthetic Pat phenotypes. Data are shown as the percentage of progeny arresting as embryos (%Emb)/percentage of progeny arresting as paralyzed, 2-fold (Pat) embryos (%Pat) (included in the Emb totals). Backgrounds have been shaded to indicate higher %Pat.

DISCUSSION

New regulatory interactions in the MS gene network

We have identified a new regulator, CEH-51, in MS specification. Our results suggest that TBX-35 and CEH-51 could participate in a ‘feed-forward’ regulatory cascade (Lee et al., 2002), in which TBX-35 activates *ceh-51*, and both TBX-35 and CEH-51 activate common target genes in MS development. There is likely to be at least one other MS lineage activator of *ceh-51* because a *ceh-51::GFP* reporter was still weakly expressed in a *tbx-35* null background (Fig. 3K). Whereas *pal-1(RNAi)* reduced coelomocyte production in *tbx-35(tm1789)* mutants (Table 1), there was no effect on *ceh-51::GFP* expression (data not shown). Instead, this activator appears to be downstream of POP-1 because simultaneous loss of *pop-1* and *tbx-35* resulted in loss of *ceh-51::GFP* expression (Fig. 3L). We also observed ectopic expression of *ceh-51* in the early E lineage in *pop-1(RNAi)* embryos (Fig. 3D,H), suggesting that POP-1 might contribute to repression of *ceh-51* in the E lineage. The observation that a *tbx-35; pop-1* background abolishes all *ceh-51::GFP* expression suggests that ectopic TBX-35 is responsible for E lineage expression of *ceh-51* in *pop-1(RNAi)*. Although we failed to detect activation of *tbx-35* in E in *pop-1(-)* embryos (Broitman-Maduro et al., 2006), such ectopic expression of *tbx-35::GFP* has been observed by others (P. Shetty and R. Lin, personal communication). We have recently shown that in the related nematode *C. briggsae*, POP-1 contributes positively to MS specification in parallel with SKN-1, and there is an apparent function for POP-1 in repression of the MS fate in E (Lin et al., 2009). Hence, these additional roles for POP-1 might be evolutionarily conserved.

Shared and distinct functions for CEH-51 and TBX-35

Although *ceh-51(tm2123); tbx-35(tm1789)* embryos have a synergistic phenotype compared with the single mutants, each gene has unique essential functions, as evidenced by their distinct

phenotypes (Figs 6 and 7). Overexpressed CEH-51 was sufficient to promote specification of muscle and coelomocytes, but was apparently not as effective at promoting pharynx development (Fig. 5), whereas overexpressed TBX-35 could specify all three tissues efficiently (Broitman-Maduro et al., 2006) (data not shown). Conversely, *ceh-51(tm2123)* mutants had only mild defects in pharynx, muscle and coelomocytes, whereas *tbx-35(tm1789)* mutants had strong defects in pharynx and muscle at 20°C (Broitman-Maduro et al., 2006) (Table 1). At 15°C, *ceh-51* is able to partially rescue these defects, resulting in a higher proportion of elongated animals and more normal specification of MS-derived tissues (Table 1). Hence, CEH-51 adds robustness to MS specification primarily at lower temperatures. In the future, identification of TBX-35 and CEH-51 target genes might explain the basis for their different activities, perhaps accounting for why CEH-51 does not rescue aspects of MS specification in *tbx-35* mutants at higher temperatures. We have identified putative TBX-35 binding sites in the promoters of *hlh-1* and *pha-4* (W.W.K.H. and M.F.M., unpublished), although we have not yet identified common targets for both TBX-35 and CEH-51.

Collaboration of T-box and NK-2 factors in mesoderm development

The apparent collaboration of TBX-35 and CEH-51 in *C. elegans* mesoderm development, downstream of MED-1,2, is highly reminiscent of the roles of related factors involved in cardiac development in other systems. In *C. elegans*, the pharynx is the structure that most closely resembles the heart, as it is a contractile pumping organ that expresses unique sets of myosins (Mango, 2007; Okkema et al., 1993). Expression of vertebrate Nkx2.5 is able to compensate for loss of *ceh-22* in the *C. elegans* pharynx, suggesting a common evolutionary origin of heart and pharynx (Haun et al., 1998). Here, we have shown that TBX-35 and CEH-51 have both distinct and shared roles in pharynx progenitor specification and development. The *Drosophila* *Nkx2.5* ortholog *tinman* is important for defining early domains that are restricted

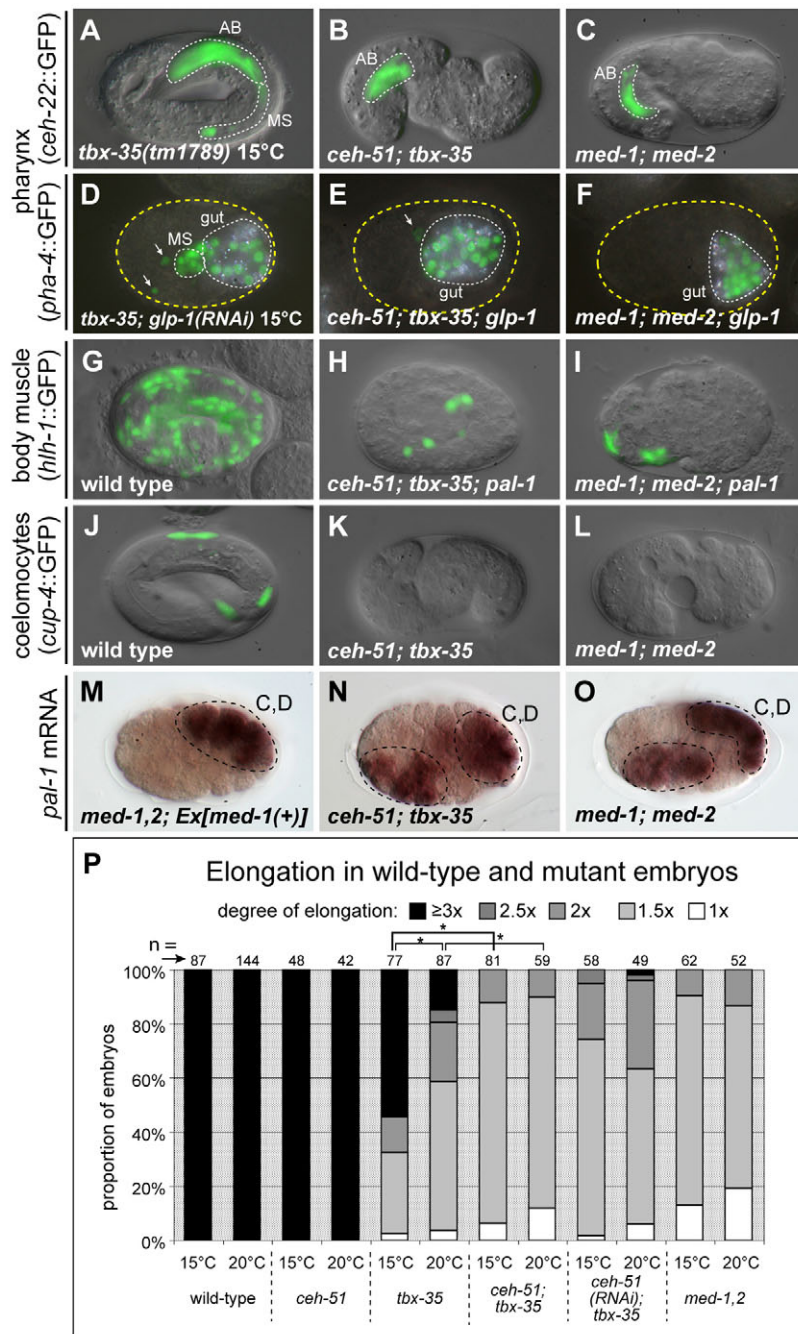


Fig. 8. Mutation of *ceh-51* and *tbx-35* together synergizes to a *med-1,2(-)* arrest phenotype.

(A-C) Pharynx muscles marked by *ceh-22::GFP* (Okkema and Fire, 1994) overlaid on DIC images. (A) Arrested 1.5-fold *tbx-35(tm1789)* embryo raised at 20°C showing AB-derived and MS-derived pharynx muscles. (B) *ceh-51(tm2123); tbx-35(tm1789)* double mutant arrested at ~1.5-fold elongation. (C) *med-1(ok804); med-2(cx9744)* double mutant. (D-F) Polarized light images to show gut granules overlaid with *pha-4::GFP* (Horner et al., 1998). (D) At 15°C, *tbx-35(tm1789); glp-1(RNAi)* embryos display 6.6 ± 0.5 ($n=23$) pharynx cells. Some additional GFP-positive cells are seen (arrows); similar 'stray' GFP expression is also seen in a *skn-1(RNAi)* background (see Table 1). Gut/rectum expression of *pha-4::GFP* coincides with birefringence of gut granules, which mark the intestine. (E) *ceh-51(tm2123); tbx-35(tm1789); glp-1(RNAi)* embryo showing a small number of pharynx cells (arrow). (F) *med-1(ok804); med-2(cx9744); glp-1(RNAi)* embryo. (G-I) Body muscle cells marked by *hlh-1::GFP* (Krause et al., 1990). (G) Wild-type embryo just before hatching. (H) *ceh-51(tm2123); pal-1(RNAi)* embryo. (I) *ceh-51(tm2123); tbx-35(tm1789); pal-1(RNAi)* embryo. (J-L) Coelomocytes marked by *cup-4::GFP* (Patton et al., 2005). (J) Wild-type embryo with four coelomocytes. (K, L) Double *ceh-51(tm2123); tbx-35(tm1789)* or *med-1(ok804); med-2(cx9744)* mutants produce little or no coelomocytes. (M) In situ hybridization showing expression of *pal-1* in the early C and D lineages (Baugh et al., 2005). (N) Ectopic expression of *pal-1* in *ceh-51(tm2123); tbx-35(tm1789)* double mutant. (O) Ectopic *pal-1* in a *med-1(ok804); med-2(cx9744)* embryo. (P) Histogram summarizing elongation of wild-type and mutant embryos. *, $P=0.05$ (χ^2 test), for some dataset pairs (comparisons among other pairs are not shown). The total number (n) of embryos scored per experiment is shown above each bar.

to forming heart, visceral muscle and some body muscles, as mutants have impairments in the development of these tissues (Azpiazu and Frasch, 1993; Bodmer, 1993). Activation of *tinman* in cardioblasts requires the T-box genes *midline* and *H15* (Reim et al., 2005). In *Xenopus*, the T-box factor Tbx5 is expressed in heart precursors and is known to be essential for heart development (Horb and Thomsen, 1999). Similarly, Nkx2.5 is expressed in early cardioblasts (Lints et al., 1993) and plays an important role in heart patterning, as *Nkx2.5* knockout mice show heart defects (Lyons et al., 1995). Finally, mouse Tbx5 and Nkx2.5 physically interact and collaborate with Gata4/5 in synergistic activation of cardiac genes (Bruneau et al., 2001; Hiroi et al., 2001; Stennard et al., 2003). Hence, the collaboration between TBX-35 and CEH-51 in *C. elegans* might be evolutionarily conserved. Future work aimed at

elucidating the gene network downstream of TBX-35 and CEH-51 might uncover further conserved aspects of cardiac and mesoderm development.

Acknowledgements

We are grateful to Shohei Mitani for *tm2123*; L. Ryan Baugh and Craig Hunter for sharing embryo transcriptome data prior to publication; Ian Hope, Michael Krause, Craig Hunter, Johnny Fares, Jeb Gaudet, Peter Okkema, Jenny Hsieh and Andy Fire for sending GFP reporter strains; Yuji Kohara for *ceh-51* ESTs; Christian Frøkjær-Jensen, Erik Jorgensen, Attila Stetak, Andy Fire and David Miller for plasmids; Serena Cervantes for preliminary characterization of the *tbx-35* elongation defects; and two anonymous reviewers for helpful comments. Some nematode strains used in this work were provided by the *Caenorhabditis* Genetics Center, which is funded by the NIH National Center for Research Resources (NCRR). This work was funded by grants from the NSF (IBN#0416922 and IOS#0643325) and NIH

(1R03HD054589-01) to M.F.M. P.W.S. is an Investigator with the Howard Hughes Medical Institute, which supported this work. Deposited in PMC for release after 6 months.

References

- Ahringer, J. (2006). Reverse genetics. In *WormBook* (ed. The C. elegans Research Community), doi:10.1895/wormbook.1.47.1, <http://www.wormbook.org>.
- Azpiaz, N. and Frasch, M. (1993). tinman and bagpipe: two homeo box genes that determine cell fates in the dorsal mesoderm of Drosophila. *Genes Dev.* **7**, 1325-1340.
- Baugh, L. R., Hill, A. A., Slonim, D. K., Brown, E. L. and Hunter, C. P. (2003). Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development* **130**, 889-900.
- Baugh, L. R., Hill, A. A., Claggett, J. M., Hill-Harfe, K., Wen, J. C., Slonim, D. K., Brown, E. L. and Hunter, C. P. (2005). The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the C. elegans embryo. *Development* **132**, 1843-1854.
- Bodmer, R. (1993). The gene tinman is required for specification of the heart and visceral muscles in Drosophila. *Development* **118**, 719-729.
- Bowerman, B., Eaton, B. A. and Priess, J. R. (1992). skn-1, a maternally expressed gene required to specify the fate of ventral blastomeres in the early C. elegans embryo. *Cell* **68**, 1061-1075.
- Bowerman, B., Draper, B. W., Mello, C. C. and Priess, J. R. (1993). The maternal gene skn-1 encodes a protein that is distributed unequally in early C. elegans embryos. *Cell* **74**, 443-452.
- Broitman-Maduro, G., Maduro, M. F. and Rothman, J. H. (2005). The noncanonical binding site of the MED-1 GATA factor defines differentially regulated target genes in the C. elegans mesoderm. *Dev. Cell* **8**, 427-433.
- Broitman-Maduro, G., Lin, K. T. H., Hung, W. and Maduro, M. (2006). Specification of the C. elegans MS blastomere by the T-box factor TBX-35. *Development* **133**, 3097-3106.
- Bruneau, B. G., Nemer, G., Schmitt, J. P., Charron, F., Robitaille, L., Caron, S., Conner, D. A., Gessler, M., Nemer, M., Seidman, C. E. et al. (2001). A murine model of Holt-Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease. *Cell* **106**, 709-721.
- Coroian, C., Broitman-Maduro, G. and Maduro, M. F. (2005). Med-type GATA factors and the evolution of mesoderm specification in nematodes. *Dev. Biol.* **289**, 444-455.
- Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806-811.
- Fox, R. M., Watson, J. D., Von Stetina, S. E., McDermott, J., Brodigan, T. M., Fukushige, T., Krause, M. and Miller, D. M., 3rd (2007). The embryonic muscle transcriptome of Caenorhabditis elegans. *Genome Biol.* **8**, R188.
- Franks, D. M., Izumikawa, T., Kitagawa, H., Sugahara, K. and Okkema, P. G. (2006). C. elegans pharyngeal morphogenesis requires both de novo synthesis of pyrimidines and synthesis of heparan sulfate proteoglycans. *Dev. Biol.* **296**, 409-420.
- Fukushige, T., Brodigan, T. M., Schrieffer, L. A., Waterston, R. H. and Krause, M. (2006). Defining the transcriptional redundancy of early bodywall muscle development in C. elegans: evidence for a unified theory of animal muscle development. *Genes Dev.* **20**, 3395-3406.
- Gaudet, J. and Mango, S. E. (2002). Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4. *Science* **295**, 821-825.
- Ghosh, T. K., Packham, E. A., Bonser, A. J., Robinson, T. E., Cross, S. J. and Brook, J. D. (2001). Characterization of the TBX5 binding site and analysis of mutations that cause Holt-Oram syndrome. *Hum. Mol. Genet.* **10**, 1983-1994.
- Goldstein, B. (1992). Induction of gut in Caenorhabditis elegans embryos. *Nature* **357**, 255-257.
- Goszczynski, B. and McGhee, J. D. (2005). Re-evaluation of the role of the med-1 and med-2 genes in specifying the C. elegans endoderm. *Genetics* **171**, 545-555.
- Hajarnavis, A., Korf, I. and Durbin, R. (2004). A probabilistic model of 3' end formation in Caenorhabditis elegans. *Nucleic Acids Res.* **32**, 3392-3399.
- Harfe, B. D. and Fire, A. (1998). Muscle and nerve-specific regulation of a novel NK-2 class homeodomain factor in Caenorhabditis elegans. *Development* **125**, 421-429.
- Harvey, R. P. (1996). NK-2 homeobox genes and heart development. *Dev. Biol.* **178**, 203-216.
- Haun, C., Alexander, J., Stainier, D. Y. and Okkema, P. G. (1998). Rescue of Caenorhabditis elegans pharyngeal development by a vertebrate heart specification gene. *Proc. Natl. Acad. Sci. USA* **95**, 5072-5075.
- Hiroi, Y., Kudoh, S., Monzen, K., Ikeda, Y., Yazaki, Y., Nagai, R. and Komuro, I. (2001). Tbx5 associates with Nkx2-5 and synergistically promotes cardiomyocyte differentiation. *Nat. Genet.* **28**, 276-280.
- Horb, M. E. and Thomsen, G. H. (1999). Tbx5 is essential for heart development. *Development* **126**, 1739-1751.
- Horner, M. A., Quintin, S., Domeier, M. E., Kimble, J., Labouesse, M. and Mango, S. E. (1998). pha-4, an HNF-3 homolog, specifies pharyngeal organ identity in Caenorhabditis elegans. *Genes Dev.* **12**, 1947-1952.
- Huang, S., Shetty, P., Robertson, S. M. and Lin, R. (2007). Binary cell fate specification during C. elegans embryogenesis driven by reiterated reciprocal asymmetry of TCF POP-1 and its coactivator beta-catenin SYS-1. *Development* **134**, 2685-2695.
- Hunt-Newbury, R., Viveiros, R., Johnsen, R., Mah, A., Anastas, D., Fang, L., Halfnight, E., Lee, D., Lin, J., Lorch, A. et al. (2007). High-throughput *in vivo* analysis of gene expression in Caenorhabditis elegans. *PLoS Biol.* **5**, e237.
- Hunter, C. P. and Kenyon, C. (1996). Spatial and temporal controls target pal-1 blastomere-specification activity to a single blastomere lineage in C. elegans embryos. *Cell* **87**, 217-226.
- Kagoshima, H., Cassata, G. and Burglin, T. R. (1999). A Caenorhabditis elegans homeobox gene expressed in the male tail, a link between pattern formation and sexual dimorphism? *Dev. Genes Evol.* **209**, 59-62.
- Kalb, J. M., Lau, K. K., Goszczynski, B., Fukushige, T., Moons, D., Okkema, P. G. and McGhee, J. D. (1998). pha-4 is Ce-fkh-1, a fork head/HNF-3alpha,beta,gamma homolog that functions in organogenesis of the C. elegans pharynx. *Development* **125**, 2171-2180.
- Kamath, R. S. and Ahringer, J. (2003). Genome-wide RNAi screening in Caenorhabditis elegans. *Methods* **30**, 313-321.
- Katic, I. and Greenwald, I. (2006). EMB-4: a predicted ATPase that facilitates lin-12 activity in Caenorhabditis elegans. *Genetics* **174**, 1907-1915.
- Kispert, A. and Herrmann, B. G. (1993). The Brachyury gene encodes a novel DNA binding protein. *EMBO J.* **12**, 4898-4899.
- Kohara, Y. (2001). Systematic analysis of gene expression of the C. elegans genome. *Tanpakushitsu Kakusan Koso* **46**, 2425-2431.
- Krause, M., Fire, A., Harrison, S. W., Priess, J. and Weintraub, H. (1990). CMyoD accumulation defines the body wall muscle cell fate during C. elegans embryogenesis. *Cell* **63**, 907-919.
- Krause, M., Harrison, S. W., Xu, S. Q., Chen, L. and Fire, A. (1994). Elements regulating cell- and stage-specific expression of the C. elegans MyoD family homolog hlh-1. *Dev. Biol.* **166**, 133-148.
- Labouesse, M. and Mango, S. E. (1999). Patterning the C. elegans embryo: moving beyond the cell lineage. *Trends Genet.* **15**, 307-313.
- Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I. et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804.
- Lei, H., Liu, J., Fukushige, T., Fire, A. and Krause, M. (2009). Caudal-like PAL-1 directly activates the bodywall muscle module regulator hlh-1 in C. elegans to initiate the embryonic muscle gene regulatory network. *Development* **136**, 1241-1249.
- Lin, K. T., Broitman-Maduro, G., Hung, W. W., Cervantes, S. and Maduro, M. F. (2009). Knockdown of SKN-1 and the Wnt effector TCF/POP-1 reveals differences in endomesoderm specification in C. briggsae as compared with C. elegans. *Dev. Biol.* **325**, 296-306.
- Lin, R. (2003). A gain-of-function mutation in oma-1, a C. elegans gene required for oocyte maturation, results in delayed degradation of maternal proteins and embryonic lethality. *Dev. Biol.* **258**, 226-239.
- Lin, R., Thompson, S. and Priess, J. R. (1995). pop-1 encodes an HMG box protein required for the specification of a mesoderm precursor in early C. elegans embryos. *Cell* **83**, 599-609.
- Lin, R., Hill, R. J. and Priess, J. R. (1998). POP-1 and anterior-posterior fate decisions in C. elegans embryos. *Cell* **92**, 229-239.
- Lints, T. J., Parsons, L. M., Hartley, L., Lyons, I. and Harvey, R. P. (1993). Nkx-2.5: a novel murine homeobox gene expressed in early heart progenitor cells and their myogenic descendants. *Development* **119**, 419-431.
- Lo, M. C., Gay, F., Odom, R., Shi, Y. and Lin, R. (2004). Phosphorylation by the beta-catenin/MAPK complex promotes 14-3-3-mediated nuclear export of TCF/POP-1 in signal-responsive cells in C. elegans. *Cell* **117**, 95-106.
- Lyons, I., Parsons, L. M., Hartley, L., Li, R., Andrews, J. E., Robb, L. and Harvey, R. P. (1995). Myogenic and morphogenetic defects in the heart tubes of murine embryos lacking the homeo box gene Nkx2-5. *Genes Dev.* **9**, 1654-1666.
- Macindoe, I., Glockner, L., Vukasin, P., Stennard, F. A., Costa, M. W., Harvey, R. P., Mackay, J. P. and Sunde, M. (2009). Conformational stability and DNA binding specificity of the cardiac T-Box transcription factor Tbx20. *J. Mol. Biol.* **389**, 606-618.
- Maduro, M. F. (2009). Structure and evolution of the C. elegans embryonic endomesoderm network. *Biochim Biophys Acta* **1789**, 250-260.
- Maduro, M. F., Meneghini, M. D., Bowerman, B., Broitman-Maduro, G. and Rothman, J. H. (2001). Restriction of mesoderm to a single blastomere by the combined action of SKN-1 and a GSK-3beta homolog is mediated by MED-1 and -2 in C. elegans. *Mol. Cell* **7**, 475-485.
- Maduro, M. F., Lin, R. and Rothman, J. H. (2002). Dynamics of a developmental switch: recursive intracellular and intranuclear redistribution of Caenorhabditis elegans POP-1 parallels Wnt-inhibited transcriptional repression. *Dev. Biol.* **248**, 128-142.

- Maduro, M. F., Kasmir, J. J., Zhu, J. and Rothman, J. H.** (2005a). The Wnt effector POP-1 and the PAL-1/Caudal homeoprotein collaborate with SKN-1 to activate *C. elegans* endoderm development. *Dev. Biol.* **285**, 510-523.
- Maduro, M., Hill, R. J., Heid, P. J., Newman-Smith, E. D., Zhu, J., Priess, J. and Rothman, J.** (2005b). Genetic redundancy in endoderm specification within the genus *Caenorhabditis*. *Dev. Biol.* **284**, 509-522.
- Maduro, M. F., Broitman-Maduro, G., Mengarelli, I. and Rothman, J. H.** (2007). Maternal deployment of the embryonic SKN-1→MED-1,2 cell specification pathway in *C. elegans*. *Dev. Biol.* **301**, 590-601.
- Mango, S. E.** (2007). The *C. elegans* pharynx: a model for organogenesis. In *WormBook* (ed. The *C. elegans* Research Community), doi:10.1895/wormbook.1.129.1, <http://www.wormbook.org>.
- Mello, C. C., Draper, B. W., Krause, M., Weintraub, H. and Priess, J. R.** (1992). The *pie-1* and *mex-1* genes and maternal control of blastomere identity in early *C. elegans* embryos. *Cell* **70**, 163-176.
- Miller, D. M., Stockdale, F. E. and Karn, J.** (1986). Immunological identification of the genes encoding the four myosin heavy chain isoforms of *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **83**, 2305-2309.
- Okkema, P. G. and Fire, A.** (1994). The *Caenorhabditis elegans* NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. *Development* **120**, 2175-2186.
- Okkema, P. G. and Krause, M.** (2005). Transcriptional regulation. In *WormBook* (ed. The *C. elegans* Research Community), doi:10.1895/wormbook.1.45.1, <http://www.wormbook.org>.
- Okkema, P. G., Harrison, S. W., Plunger, V., Aryana, A. and Fire, A.** (1993). Sequence requirements for myosin gene expression and regulation in *Caenorhabditis elegans*. *Genetics* **135**, 385-404.
- Page, B. D., Diede, S. J., Tenlen, J. R. and Ferguson, E. L.** (2007). EEL-1, a Hect E3 ubiquitin ligase, controls asymmetry and persistence of the SKN-1 transcription factor in the early *C. elegans* embryo. *Development* **134**, 2303-2314.
- Patton, A., Knuth, S., Schaheen, B., Dang, H., Greenwald, I. and Fares, H.** (2005). Endocytosis function of a ligand-gated ion channel homolog in *Caenorhabditis elegans*. *Curr. Biol.* **15**, 1045-1050.
- Phillips, B. T., Kidd, A. R., 3rd, King, R., Hardin, J. and Kimble, J.** (2007). Reciprocal asymmetry of SYS-1/beta-catenin and POP-1/TCF controls asymmetric divisions in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **104**, 3231-3236.
- Priess, J. R., Schnabel, H. and Schnabel, R.** (1987). The *glp-1* locus and cellular interactions in early *C. elegans* embryos. *Cell* **51**, 601-611.
- Reece-Hoyes, J. S., Shingles, J., Dupuy, D., Grove, C. A., Walhout, A. J., Vidal, M. and Hope, I. A.** (2007). Insight into transcription factor gene duplication from *Caenorhabditis elegans* Promoterome-driven expression patterns. *BMC Genomics* **8**, 27.
- Reim, I., Mohler, J. P. and Frasch, M.** (2005). Tbx20-related genes, mid and H15, are required for tinman expression, proper patterning, and normal differentiation of cardioblasts in *Drosophila*. *Mech. Dev.* **122**, 1056-1069.
- Rochelleau, C. E., Downs, W. D., Lin, R., Wittmann, C., Bei, Y., Cha, Y. H., Ali, M., Priess, J. R. and Mello, C. C.** (1997). Wnt signaling and an APC-related gene specify endoderm in early *C. elegans* embryos. *Cell* **90**, 707-716.
- Rochelleau, C. E., Yasuda, J., Shin, T. H., Lin, R., Sawa, H., Okano, H., Priess, J. R., Davis, R. J. and Mello, C. C.** (1999). WRM-1 activates the LIT-1 protein kinase to transduce anterior/posterior polarity signals in *C. elegans*. *Cell* **97**, 717-726.
- Shaham, S.** (2006). Methods in cell biology. In *WormBook* (ed. The *C. elegans* Research Community), doi:10.1895/wormbook.1.49.1, <http://www.wormbook.org>.
- Shelton, C. A. and Bowerman, B.** (1996). Time-dependent responses to *glp-1*-mediated inductions in early *C. elegans* embryos. *Development* **122**, 2043-2050.
- Shetty, P., Lo, M. C., Robertson, S. M. and Lin, R.** (2005). *C. elegans* TCF protein, POP-1, converts from repressor to activator as a result of Wnt-induced lowering of nuclear levels. *Dev. Biol.* **285**, 584-592.
- Shirayama, M., Soto, M. C., Ishidate, T., Kim, S., Nakamura, K., Bei, Y., van den Heuvel, S. and Mello, C. C.** (2006). The conserved kinases CDK-1, GSK-3, KIN-19, and MBK-2 promote OMA-1 destruction to regulate the oocyte-to-embryo transition in *C. elegans*. *Curr. Biol.* **16**, 47-55.
- Stennard, F. A., Costa, M. W., Elliott, D. A., Rankin, S., Haast, S. J., Lai, D., McDonald, L. P., Niederreither, K., Dolle, P., Bruneau, B. G. et al.** (2003). Cardiac T-box factor Tbx20 directly interacts with Nkx2-5, GATA4, and GATA5 in regulation of gene expression in the developing heart. *Dev. Biol.* **262**, 206-224.
- Sulston, J. E. and Hodgkin, J.** (1988). Methods. In *The Nematode Caenorhabditis elegans* (ed. W. B. Wood), pp. 587-606. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
- Sulston, J. E., Schierenberg, E., White, J. G. and Thomson, J. N.** (1983). The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**, 64-119.
- Syntichaki, P. and Tavernarakis, N.** (2004). Genetic models of mechanotransduction: the nematode *Caenorhabditis elegans*. *Physiol. Rev.* **84**, 1097-1153.
- Thorpe, C. J., Schlesinger, A., Carter, J. C. and Bowerman, B.** (1997). Wnt signaling polarizes an early *C. elegans* blastomere to distinguish endoderm from mesoderm. *Cell* **90**, 695-705.
- Williams, B. D. and Waterston, R. H.** (1994). Genes critical for muscle development and function in *Caenorhabditis elegans* identified through lethal mutations. *J. Cell Biol.* **124**, 475-490.

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 $\Delta pes-10$ 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 motoneurons.

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 resembles 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/Explainers 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 $\Delta pes-10::4X-NLS::eGFP::LacZ::unc-54$ 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 $\Delta pes-10::4X-NLS::YFP::unc-54$ sequence. Control un-mutated and sequenced enhancers were fused to pPD122.53 with CFP replacing GFP, to give a $\Delta pes-10::4X-NLS::CFP::unc-54$ 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_JD02.0 05	299	WBGene00016441 C35D10.3 (1 elegans, 3 briggsae, 10 remanei, 6 brenneri).
		Cbre_JD02.0 06	509	WBGene00016441 C35D10.3 (1 elegans, 3 briggsae, 10 remanei, 6 brenneri).
		Cbre_JD02.0 07	314	WBGene00016441 C35D10.3 (1 elegans, 3 briggsae, 10 remanei, 6 brenneri).
		Cbre_JD02.0 08	851	WBGene00016441 C35D10.3 (1 elegans, 3 briggsae, 10 remanei, 6 brenneri).
		Cbre_JD02.0 09	316	WBGene00016441 C35D10.3 (1 elegans, 3 briggsae, 10 remanei, 6 brenneri).
		Cbre_JD02.0 10	98	
Cbre_JD0 3	16,00 3	Cbre_JD03.0 01	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_JD03.0 02	120	
		Cbre_JD03.0 03	221	(1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD03.0 04	46	
Cbre_JD0 4	20,54 6	Cbre_JD04.0 01	403	WBGene00020867 shc-2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD04.0 02	601	WBGene00020868 T27F7.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD04.0 03	121	WBGene00020866 T27F7.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD04.0 04	127	WBGene00003425 msp-10, WBGene00003432 msp-36, WBGene00003449 msp-56, and WBGene00003463 msp-76 (4 elegans, 3 briggsae, 16 remanei, 1 brenneri).
		Cbre_JD04.0 05	52	
		Cbre_JD04.0 06	164	WBGene00004382 rnh-1.0 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD04.0 07	180	WBGene00004382 rnh-1.0 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD04.0 08	73	
		Cbre_JD04.0 09	70	

		Cbre_JD04.0 10	247	WBGene00007303 rnh-1.3 [*] (1 elegans, 1 briggsae, 1 brenneri).
Cbre_JD0 5	10,51 4	Cbre_JD05.0 01	81	(2 brenneri).
		Cbre_JD05.0 02	127	(2 brenneri).
		Cbre_JD05.0 03	1,331	WBGene00021678 Y48G1C.5 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri).
		Cbre_JD05.0 04	115	WBGene00003097 lys-8 [*] (1 elegans, 1 briggsae, 2 remanei, 2 brenneri).
Cbre_JD0 6	18,12 0	Cbre_JD06.0 01	676	WBGene00020183 T03D3.5 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri).
		Cbre_JD06.0 02	324	WBGene00017090 E01A2.8 and WBGene00044697 K05F6.11 (2 elegans, 2 briggsae, 1 remanei, 2 brenneri).
		Cbre_JD06.0 03	231	
		Cbre_JD06.0 04	284	
Cbre_JD0 7	66,84 9	Cbre_JD07.0 01	1,272	WBGene00000549 cls-2 and WBGene00015580 C07H6.3 (2 elegans, 2 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD07.0 02	912	WBGene00000537 clk-2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD07.0 03	513	WBGene00000854 cux-7 [*] (1 elegans, 1 briggsae, 1 brenneri).
		Cbre_JD07.0 04	105	WBGene00015579 C07H6.2 [*] (1 elegans, 1 briggsae, 1 brenneri, 1 ps1010).
		Cbre_JD07.0 05	703	WBGene00002986 lig-4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD07.0 06	95	
		Cbre_JD07.0 07	252	WBGene00003024 lin-39 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD07.0 08	78	
		Cbre_JD07.0 09	42	
		Cbre_JD07.0 10	68	
		Cbre_JD07.0 11	54	
		Cbre_JD07.0 12	202	WBGene00000437 ceh-13 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD07.0 13	139	(2 brenneri).

		Cbre_JD07.0 14	141	(2 brenneri).
		Cbre_JD07.0 15	260	WBGene00022102 Y69F12A.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
Cbre_JD0 8	27,63 4	Cbre_JD08.0 01	341	WBGene00013956 ZK265.3 [*] (1 elegans, 1 remanei, 1 brenneri).
		Cbre_JD08.0 02	393	WBGene00000639 col-63 [*] (1 elegans, 2 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD08.0 03	290	WBGene00000433 ceh-8 [*] (1 elegans, 1 remanei, 1 brenneri).
		Cbre_JD08.0 04	283	WBGene00044094 ZK265.9 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD08.0 05	189	WBGene00013958 ZK265.6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD08.0 06	414	WBGene00013957 sre-23 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD08.0 07	43	
		Cbre_JD08.0 08	370	WBGene00013959 ZK265.7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
Cbre_JD0 9	32,96 8	Cbre_JD09.0 01	326	WBGene00000603 col-14 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD09.0 02	1,255	WBGene00011530 T06D8.10, WBGene00016700 C46A5.4, and WBGene00019613 K10B4.1 (3 elegans, 3 briggsae, 4 remanei, 1 brenneri).
		Cbre_JD09.0 03	479	WBGene00016848 C50F7.10 and WBGene00017103 E02H9.5 (2 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD09.0 04	417	WBGene00016842 C50F7.1 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD09.0 05	373	WBGene00011290 R102.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD09.0 06	149	WBGene00011291 R102.4 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD09.0 07	184	
		Cbre_JD09.0 08	266	WBGene00021541 Y42H9B.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD09.0 09	318	WBGene00016130 C26B2.8 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD09.0 10	136	WBGene00016129 C26B2.7 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD09.0 11	335	WBGene00016128 C26B2.6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD09.0	862	WBGene00016124 C26B2.1 [*] (1 elegans, 1

		12		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_JD11.0 10	287	WBGene00001210 egl-46 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD11.0 11	76	
		Cbre_JD11.0 12	84	
		Cbre_JD11.0 13	419	WBGene00019655 K11G9.5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD11.0 14	75	WBGene00003473 mtl-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD11.0 15	80	
		Cbre_JD11.0 16	71	WBGene00020947 W02F12.2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
Cbre_JD1 2	36,17 8	Cbre_JD12.0 01	304	WBGene00001668 gpa-6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD12.0 02	668	WBGene00009844 cwp-5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD12.0 03	90	WBGene00003741 nlp-3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD12.0 04	36	
		Cbre_JD12.0 05	60	
		Cbre_JD12.0 06	80	(1 briggsae, 1 remanei, 1 brenneri).
Cbre_JD1 3	48,93 4	Cbre_JD13.0 01	261	WBGene00013891 ZC434.3 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri, 2 ps1010).
		Cbre_JD13.0 02	203	WBGene00013891 ZC434.3 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri, 2 ps1010).
		Cbre_JD13.0 03	884	WBGene00002153 irs-2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD13.0 04	122	(1 briggsae, 1 brenneri).
		Cbre_JD13.0 05	136	WBGene00007708 C25A1.6 [*] (1 elegans, 1 remanei, 1 brenneri).
		Cbre_JD13.0 06	316	WBGene00007707 C25A1.5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD13.0 07	449	WBGene00007706 C25A1.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD13.0 08	193	WBGene00001442 fkh-10 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).

		Cbre_JD13.0 09	225	WBGene00007705 C25A1.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD13.0 10	372	WBGene00006447 tag-72 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD13.0 11	812	WBGene00002994 lin-5 and WBGene00008508 F01G10.5 (2 elegans, 1 briggsae, 4 remanei, 1 brenneri).
		Cbre_JD13.0 12	369	WBGene00003000 lin-11 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD13.0 13	642	WBGene00013860 ZC247.2 and WBGene00013895 ZC434.9 (2 elegans, 2 briggsae, 2 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD13.0 14	1,747	WBGene00013859 ZC247.1 (1 elegans, 6 briggsae, 18 remanei, 1 brenneri).
Cbre_JD1 4	34,73 8	Cbre_JD14.0 01	139	WBGene00001426 fkb-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD14.0 02	1,432	WBGene00006490 tag-144 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD14.0 03	75	WBGene00009496 F36H1.11 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD14.0 04	117	WBGene00009497 F36H1.12 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD14.0 05	483	WBGene00002992 lin-3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Cbre_JD14.0 06	132	WBGene00012382 Y5F2A.1 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD14.0 07	131	WBGene00012383 Y5F2A.2 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD14.0 08	78	
		Cbre_JD14.0 09	450	WBGene00012385 Y5F2A.4 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD14.0 10	645	WBGene00010882 atgr-7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD14.0 11	147	WBGene00002344 let-70 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD14.0 12	541	WBGene00000246 bcc-1 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD14.0 13	214	WBGene00010883 M7.7 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri).
Cbre_JD1 5	13,75 1	Cbre_JD15.0 01	204	WBGene00018965 F56D2.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD15.0 02	420	WBGene00022632 ZC581.2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD15.0 03	123	WBGene00017299 F09F7.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).

Cbre_JD1 6	19,58 6	Cbre_JD16.0 01	152	WBGene00003371 mlc-3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD16.0 02	1,152	WBGene00016140 rpb-2 and WBGene00017300 F09F7.3 (2 elegans, 2 briggsae, 2 remanei, 1 brenneri).
		Cbre_JD16.0 03	386	WBGene00017301 F09F7.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD16.0 04	314	WBGene00017304 F09F7.7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD16.0 05	87	WBGene00017305 nspb-12 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD16.0 06	74	
Cbre_JD1 7	35,36 2	Cbre_JD17.0 01	208	(2 brenneri).
		Cbre_JD17.0 02	318	(2 brenneri).
		Cbre_JD17.0 03	383	WBGene00008401 D2005.6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD17.0 04	15	
		Cbre_JD17.0 05	170	WBGene00003746 nlp-8 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
Cbre_JD1 8	6,580	Cbre_JD18.0 01	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_JD18.0 02	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_JD1	25,57	Cbre_JD19.0	2,149	(1 remanei, 1 brenneri).

9	8	01		
		Cbre_JD19.0 02	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_JD19.0 03	1,415	WBGene00004323 rde-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD19.0 04	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_JD2 0	38,44 1	Cbre_JD20.0 01	468	WBGene00011041 R05H5.7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD20.0 02	149	WBGene00011038 R05H5.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD20.0 03	435	WBGene00011039 R05H5.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD20.0 04	238	WBGene00011040 R05H5.5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD20.0 05	49	
		Cbre_JD20.0 06	516	WBGene00011331 T01E8.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD20.0 07	68	
		Cbre_JD20.0 08	386	WBGene00004334 ref-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
Cbre_JD2 1	33,64 8	Cbre_JD21.0 01	81	(2 brenneri).
		Cbre_JD21.0 02	127	(2 brenneri).
		Cbre_JD21.0	1,331	WBGene00021678 Y48G1C.5 [*] (1 elegans, 1

		03		briggsae, 1 remanei, 2 brenneri).
		Cbre_JD21.0 04	154	WBGene00003097 lys-8 [*] (1 elegans, 1 briggsae, 2 remanei, 2 brenneri).
		Cbre_JD21.0 05	596	WBGene00020183 T03D3.5 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri).
		Cbre_JD21.0 06	366	WBGene00017090 E01A2.8 and WBGene00044697 K05F6.11 (2 elegans, 2 briggsae, 1 remanei, 2 brenneri).
		Cbre_JD21.0 07	371	WBGene00010366 H05L14.1 (1 elegans, 2 briggsae, 3 remanei, 1 brenneri).
		Cbre_JD21.0 08	381	WBGene00005749 srw-2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri).
		Cbre_JD21.0 09	326	WBGene00008568 F08A8.5 and WBGene00012070 T26H5.8 (2 elegans, 1 briggsae, 2 remanei, 1 brenneri).
Cbre_JD2 2	33,58 9	Cbre_JD22.0 01	427	
		Cbre_JD22.0 02	73	
		Cbre_JD22.0 03	156	(7 briggsae, 1 brenneri).
		Cbre_JD22.0 04	118	(4 remanei, 1 brenneri).
		Cbre_JD22.0 05	67	
		Cbre_JD22.0 06	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_JD0 1	43,54 4	Csp3_JD01.0 01	975	WBGene00018721 polh-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD01.0 02	578	WBGene00004491 rps-22 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD01.0 03	383	WBGene00017732 F23C8.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD01.0	4,291	WBGene00000396 cdh-4 [*] (1 elegans, 1

		04		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

		04		briggsae, 1 remanei, 1 ps1010).
		Csp3_JD03.0 05	181	WBGene00001309 emr-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD03.0 06	457	WBGene00006461 tag-96 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD03.0 07	317	WBGene00004743 scm-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD03.0 08	872	WBGene00022139 tag-305 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD03.0 09	432	WBGene00001007 dli-1 [*] (1 elegans, 1 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD03.0 10	361	WBGene00009140 F26A3.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
Csp3_JD0 4	81,32 8	Csp3_JD04.0 01	503	WBGene00000117 alh-11 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD04.0 02	477	WBGene00001573 gei-16 (1 elegans, 1 ps1010).
		Csp3_JD04.0 03	949	WBGene00001573 gei-16 (1 elegans, 1 ps1010).
		Csp3_JD04.0 04	181	
		Csp3_JD04.0 05	1,332	WBGene00020550 T17H7.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD04.0 06	167	
		Csp3_JD04.0 07	177	
		Csp3_JD04.0 08	191	WBGene00003102 mab-5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD04.0 09	252	WBGene00015591 C08C3.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD04.0 10	211	WBGene00001174 egl-5 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD04.0 11	1,086	WBGene00000768 cor-1 and WBGene00007983 C36E8.4 (2 elegans, 2 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD04.0 12	775	
		Csp3_JD04.0 13	340	WBGene00003162 mdh-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD04.0 14	117	WBGene00019509 K07H8.9 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
Csp3_JD0 5	66,53 5	Csp3_JD05.0 01	213	WBGene00004418 rpl-7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD05.0 02	251	WBGene00018774 F53G12.9 [*] (1 elegans, 1 remanei, 1 ps1010).

		Csp3_JD05.0 03	1,639	WBGene00003210 mel-28 (1 elegans, 2 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD05.0 04	1,876	WBGene00002040 hum-7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD05.0 05	503	WBGene00022709 ZK354.8 [*] (1 elegans, 1 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD05.0 06	291	WBGene00014083 ZK795.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD05.0 07	1,195	WBGene00006961 xnp-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD05.0 08	304	WBGene00012156 ebp-2 [*] (1 elegans, 2 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD05.0 09	347	WBGene00006447 tag-72 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD05.0 10	344	WBGene00003000 lin-11 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD05.0 11	1,077	WBGene00013860 ZC247.2 and WBGene00013895 ZC434.9 (2 elegans, 2 briggsae, 2 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD05.0 12	335	WBGene00011340 ugt-30, WBGene00015693 ugt-28, and WBGene00021709 ugt-29 (3 elegans, 1 briggsae, 2 remanei, 2 ps1010).
		Csp3_JD05.0 13	332	WBGene00011340 ugt-30, WBGene00015693 ugt-28, and WBGene00021709 ugt-29 (3 elegans, 1 briggsae, 2 remanei, 2 ps1010).
		Csp3_JD05.0 14	295	WBGene00013893 ZC434.7 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD05.0 15	82	
		Csp3_JD05.0 16	1,841	WBGene00000148 aph-2 and WBGene00001337 ers-2 (2 elegans, 2 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD05.0 17	258	WBGene00013891 ZC434.3 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri, 2 ps1010).
		Csp3_JD05.0 18	276	WBGene00013891 ZC434.3 [*] (1 elegans, 1 briggsae, 1 remanei, 2 brenneri, 2 ps1010).
		Csp3_JD05.0 19	271	WBGene00013892 ZC434.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
Csp3_JD0 6	60,75 7	Csp3_JD06.0 01	486	WBGene00005663 srs-2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD06.0 02	301	WBGene00008147 C47E12.2 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD06.0 03	546	WBGene00008148 C47E12.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).

		Csp3_JD06.0 04	325	WBGene00022707 ZK354.6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD06.0 05	441	WBGene00009686 F44D12.9 (1 elegans, 2 briggsae, 3 remanei, 1 ps1010).
		Csp3_JD06.0 06	528	WBGene00003992 ppl-1 and WBGene00003994 ppl-3 (2 elegans, 1 briggsae, 1 ps1010).
		Csp3_JD06.0 07	234	WBGene00011746 T13F2.6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD06.0 08	223	WBGene00002274 lec-11 [*] (1 elegans, 1 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD06.0 09	83	
		Csp3_JD06.0 10	510	WBGene00003603 nhr-4 [*] (1 elegans, 1 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD06.0 11	391	WBGene00002992 lin-3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD06.0 12	136	WBGene00009497 F36H1.12 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD06.0 13	1,476	WBGene00006490 tag-144 [*] (1 elegans, 1 briggsae, 2 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD06.0 14	271	WBGene00001426 fkb-1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 brenneri, 1 ps1010).
		Csp3_JD06.0 15	858	WBGene00015571 C07G1.2 [*] (1 elegans, 1 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD06.0 16	641	WBGene00003838 ocr-1, WBGene00003839 ocr-2, and WBGene00003840 ocr-3 (3 elegans, 3 briggsae, 3 remanei, 1 ps1010).
Csp3_JD0 7	30,01 2	Csp3_JD07.0 01	245	WBGene00015156 B0361.2 [*] (1 elegans, 1 briggsae, 2 remanei, 1 ps1010).
		Csp3_JD07.0 02	681	WBGene00004905 snf-6 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD07.0 03	351	WBGene00019716 M01G5.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD07.0 04	27	
		Csp3_JD07.0 05	138	
		Csp3_JD07.0 06	849	WBGene00019715 M01G5.1 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD07.0 07	340	WBGene00022793 ZK686.3 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD07.0 08	218	WBGene00022794 ZK686.4 [*] (1 elegans, 1 briggsae, 1 remanei, 1 ps1010).
		Csp3_JD07.0	657	WBGene00008167 C48B4.1,

		09		WBGene00008564 F08A8.1, WBGene00008565 F08A8.2, WBGene00008566 F08A8.3, and WBGene00008567 F08A8.4 (5 <i>elegans</i> , 5 <i>briggsae</i> , 4 <i>remanei</i> , 1 <i>ps1010</i>).
--	--	----	--	--

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 <i>ceh-13</i>	3' stop with respect to <i>ceh-13</i>	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..753769 4
I4	-16593	-12411	III:7538991..754317 3
I5	-11977	-11697	III:7543607..754388 7
I6	-11106	-10890	III:7544478..754469 4
I7	-10195	-6925	III:7545389..754865 9
I8	-5805	-2899	III:7549779..755268 5
I9	-1783	-826	III:7553801..755475 8
W2	-11697	-5805	III:7543887..754977 9

B. Coordinates of MUSSA matches in initial study

Element	5' start with respect to <i>ceh-13</i>	3' stop with respect to <i>ceh-13</i>	Chromosomal location
N1	-24807	-24783	III:7530777..753080 1
	-24762	-24735	III:7530822..753084 9
	-24677	-24629	III:7530907..753095 5
	-24060	-24040	III:7531524..753154 4
	-24030	-24006	III:7531554..753157 8
N2	-23499	-23450	III:7532085..753213 4
	-23365	-23339	III:7532219..753224 5
N3	-22460	-22433	III:7533124..753315 1
N4	-18832	-18815	III:7536752..753676 9
	-18802	-18769	III:7536782..753681 5
	-18742	-18719	III:7536842..753686 5
N5	-17606	-17578	III:7537978..753800 6
N6	-12362	-12338	III:7543222..754324 6

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 <i>ceh-13</i>	3' end with respect to <i>ceh-13</i>	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	CAAGGCCTGCAGGCATGCAAGCCCATAACCGAAGCAATTCTCTC A
N1R_XbaI	ATATCTAGATGTTACACCGTGTCTCCCTCAT
N1L_HinDIII	TCAAAAAGCTTCCATAACCGAAGCAATTCTCTCA
N2L_fus	CAAGGCCTGCAGGCATGCAAGCTTTTAAGCGTCTGCGTCTGAAGT
N2R_XbaI	ATATCTAGATCTCCACTGAATATCGCCAGTTC
N2L_HinDIII	TCAAAAAGCTTTTTTAAGCGTCTGCGTCTGAAGT
N3L_fus	CAAGGCCTGCAGGCATGCAAGCGCACCCCTAGATCAACAAGCTTC A
N3R_XbaI	ATATCTAGATTTGGCAAAACAATGGTCTCAC
N3L_StuI	TCAAAGGCCTGCACCCTAGATCAACAAGCTTCA
N4L_fus	CAAGGCCTGCAGGCATGCAAGCTTAAACGTTTTCTGCCACAAAG G
N4R_StuI	TCAAAGGCCTTTTTGTTTCTAAAAGCGGCAACT
N5L_fus	CAAGGCCTGCAGGCATGCAAGCCAAATTCTCAGAGCCACAACAC A
N5R_SphI	GCTGCATGCTACCCCTGTGCAACTCAACAAT
N6L_fus	CAAGGCCTGCAGGCATGCAAGCAGCCAAATGAAGTGCCAATTTT A
N6R_HinDIII	TTACAAGCTTGCCCATCTTCGAAAATTTTGTTT

N7L_fus	CAAGGCCTGCAGGCATGCAAGCTTTTTCTTATTTAACCTGCACCA CA
N7L_HinDIII	TCAAAAAGCTTGGAATGTCGGAGTCCAAAAGAT
N7R_XbaI	ATATCTAGAGGAATGTCGGAGTCCAAAAGAT
N8L_SalI	CATTAGTCGACACAACCTTTCGCCTGTGTCTGTTT
N8R_fus	CAAGGCCTGCAGGCATGCAAGCCCCTCTAGACACCTGTTGTTCTT CT
N9L_StuI	TCAAAAAGGCCTTTTCAAAAAGTCGCCTTTACAGTCA
N9R_fus	CAAGGCCTGCAGGCATGCAAGCCCCGATTAAAAGTTGTAAGGCA AT
N10L_StuI	TCAAAAAGGCCTACTGTAGCCCCGACACTGATGTTC
N10R_fus	CAAGGCCTGCAGGCATGCAAGCCTATGAGGAGATGGACACGGAG T
N11L_HinDIII	TCAAAAAGCTTCTCCTTCTTTTCCCCGTGTCC
N11R_fus	CAAGGCCTGCAGGCATGCAAGCAGTGGAGCTCATGCTGGAAAAT A
I0L_fus	CAAGGCCTGCAGGCATGCAAGCTATGCTGTTTCGTTGTCGCTTCT
I0R	TGAGAGAATTGCTTCGGTTATGG
I1L_fus	CAAGGCCTGCAGGCATGCAAGCATGAGGGAGAACACGGTGTAAC A
I1R	ACTTCAGACGCAGACGCTTAAAA
I2L_fus	CAAGGCCTGCAGGCATGCAAGCGAACTGGCGATATTCAGTGGAG A
I2R	TGAAGCTTGTTGATCTAGGGTGC
I3L_fus	CAAGGCCTGCAGGCATGCAAGCAGTTGCCGCTTTTAGGAACAAA A
I3R	TGTGTTGTGGCTCTGAGAATTTG
I4L_fus	CAAGGCCTGCAGGCATGCAAGCATTGTTGAGTTGCACAGGGGT A
I4R	TAAAATTGGCACTTCATTTGGCT
I5L_fus	CAAGGCCTGCAGGCATGCAAGCAAACAAAATTTTCGAAGATGGG C
I5R	TGTGGTGCAGGTAAATAAGAAAAA
I6L	ATCTTTTGGACTCCGACATTCC
I6R_fus	CAAGGCCTGCAGGCATGCAAGCAAACAGACACAGGCGAAAGTTG T
I7L	AGAAGAACAACAGGTGTCTAGAGGG
I7R_fus	CAAGGCCTGCAGGCATGCAAGCTGACTGTAAAGGCGACTTTTGA AA
I8L	ATTGCCTTACAACCTTTAATCGGG
I8R_fus	CAAGGCCTGCAGGCATGCAAGCGAACATCAGTGTCGGGCTACAG T
I9L	ACTCCGTGTCCATCTCCTCATAG
I9R_fus	CAAGGCCTGCAGGCATGCAAGCGGACACGGGGAAAAGAAGGAG
N1mL	TACCGCTGCGGGGAACAGTTTCATAAACCTGAGTTGCTCTGATAGCTG

	TGATG
N1mR	CATCACAGCTATCAGAGCAACTCAGGTTTATGAACTGTTCCCCGCA GCGGTA
N2-1mL	GAAAGTGAGTGGCGGGGAGCACAGTTCTGGAAGATAAATGGGCTCG CGAC
N2-1mR	GTCGCGAGCCATTTATCTTCCAGAACTGTGCTCCCCGCCACTCACTT TC
N2-2mL	GCGTCGCCTTCTTCCTTTAGTAAACTGTACTTCGTAGTGGAGAGAGG GAAAAGAAG
N2-2mR	CTTCTTTTCCCTCTCTCCACTACGAAGTACAGTTTTACTAAAGGAAGA AGGCGACGC
N3mL	GAGACAAACAGCGGGAATCAAAGTTCTAATTAACCTTCTCTCACTCT TTCCTCTC
N3mR	GAGAGTGAAAGAGTGAGAGGAAGGTTAATTAGAACTTTGATTCCCGC TGTTTGTCTC
N7mL	AAAAGAGGGTAAAGATTTCTAAATACCCACGGTAATTCAACTCTCAC CAGACGTACG
N7mR	GTCTGGTGAGAGTTGAATTACCGTGGGTATTTAGAAATCTTTACCTC TTTCCATC
MmN3L_XbaI	ACATATCTAGATGTTTGCCTCCTGATCTGC
MmN3R_Hin DIII	TCAAAAAGCTTGAAGTTGATGGCGAAGGAAG
MmN3L_fusio n	CAAGGCCTGCAGGCATGCAAGCTGTTTGCCTCCTGATCTGC
MmN7L_Hin DIII	TCAAAAAGCTTGCCTGGAGGAGTCCTAACC
MmN7R_XbaI	ACATATCTAGAACTCCCTTCGACTCCATCTG
MmN7R_fusio n	CAAGGCCTGCAGGCATGCAAGCACTCCCTTCGACTCCATCTG
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
<i>lin-3</i>	<i>elegans</i>	IV:11053607..11063483
	<i>briggsae</i>	chrIV:5701665..5708512 [antisense]
	<i>remanei</i>	Supercontig32:284661..291046

	<i>brenneri</i>	CB5161_lin-3.tfa:12411..19047
	sp. 3 PS1010	PS1010_lin-3.tfa:31409..36034 [antisense]
ACEL	<i>elegans</i>	IV:11059133..11059192
	<i>briggsae</i>	chrIV:5704301..5704360 [antisense]
	<i>remanei</i>	Contig32.18:21275..21334
	<i>brenneri</i>	CB5161_lin-3.tfa:16249..16308
	sp. 3 PS1010	n/a [5' flank was PS1010_lin-3.tfa:34099..36034; antisense]
<i>lin-11</i>	<i>elegans</i>	I:10241073..10255621
	<i>briggsae</i>	chrI:6218293..6230072 [antisense]
	<i>remanei</i>	Supercontig31:626189..635406
	<i>brenneri</i>	CB5161_lin-11.tfa:26842..36289
	sp. 3 PS1010	PS1010_lin-11.tfa:31373..37085
uterine	<i>elegans</i>	I:10245795..10246254
	<i>briggsae</i>	chrI:6225822..6226281 [antisense]
	<i>remanei</i>	Contig31.36:12788..13247
	<i>brenneri</i>	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
CACCT G	E-box (<i>lin-3</i>)	24.52 [1]	30.04 [1]	30.04 [1]	34.68 [1]	12.23 [1]
ACCCT G	Ftz-F1 (<i>lin-3</i>)	15.72 [2]	19.25 [2]	19.25 [2]	22.23 [2]	8.67 [2]
ATGGG A	LAG-1 (<i>lin-11</i>)	[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; Gribkov 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*.