2007

Identification of muscle-specific regulatory modules in Caenorhabditis elegans

Guoyan Zhao
Washington University School of Medicine in St. Louis

Lawrence A. Schriefer
Washington University School of Medicine in St. Louis

Gary D. Stormo
Washington University School of Medicine in St. Louis

Identification of muscle-specific regulatory modules in *Caenorhabditis elegans*

Guoyan Zhao, Lawrence A. Schriefer and Gary D. Stormo

Genome Res. 2007 17: 348-357 originally published online February 6, 2007
Access the most recent version at doi:10.1101/gr.5989907

Methods

Identification of muscle-specific regulatory modules in *Caenorhabditis elegans*

Guoyan Zhao, Lawrence A. Schriefer, and Gary D. Stormo[1]

Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110, USA

Transcriptional regulation is the major regulatory mechanism that controls the spatial and temporal expression of genes during development. This is carried out by transcription factors (TFs), which recognize and bind to their cognate binding sites. Recent studies suggest a modular organization of TF-binding sites, in which clusters of transcription-factor binding sites cooperate in the regulation of downstream gene expression. In this study, we report our computational identification and experimental verification of muscle-specific *cis*-regulatory modules in *Caenorhabditis elegans*. We first identified a set of motifs that are correlated with muscle-specific gene expression. We then predicted muscle-specific regulatory modules based on clusters of those motifs with characteristics similar to a collection of well-studied modules in other species. The method correctly identifies 88% of the experimentally characterized modules with a positive predictive value of at least 65%. The prediction accuracy of muscle-specific expression on an independent test set is highly significant ($P < 0.0001$). We performed in vivo experimental tests of 12 predicted modules, and 10 of those drive muscle-specific gene expression. These results suggest that our method is highly accurate in identifying functional sequences important for muscle-specific gene expression and is a valuable tool for guiding experimental designs.

[Supplemental material is available online at www.genome.org.]

In metazoans, the gene-regulatory information that directs development is encoded in their genomic DNA sequence. The temporal and spatial expression pattern of genes is controlled by short *cis*-regulatory elements that act as binding sites for transcription factors. Through interactions with the basal transcription apparatus and other regulatory proteins, transcription factors determine either activation or repression of the target gene at a particular developmental time or within a particular cell or tissue. Therefore, identification of *cis*-regulatory elements and their binding proteins constitute an important part of deciphering the role of noncoding sequences. However, the individual binding of a transcription factor to a regulatory element is rarely sufficient to confer context-specific expression. Mounting evidence suggests that complex, cooperative protein–protein interactions between transcription factors are required to determine gene expression patterns (Arnone and Davidson 1997; Kamachi et al. 2000; Li et al. 2000; Remenyi et al. 2004). Therefore, identification of all of the component regulatory elements and understanding how they interact with each other are crucial to fully understanding the transcriptional regulatory network.

Given the fast increasing number of genome sequences, our ability to decipher the encoded information lags far behind. For example, *Caenorhabditis elegans* is the first metazoan organism whose genome was sequenced. However, our understanding of the sequences that control tissue-specific gene expression is still limited. This limited understanding comes mainly from experimental investigation of the regulatory sequences of individual genes, which began almost 20 yr ago (Spieth et al. 1988). In *C. elegans*, *cis*-regulation of tissue-specific gene expression is known only for a few genes in some tissues, such as hypodermal cell, excretory cell, vulva, muscle, and neurons (Okkema et al. 1993; Gilleard et al. 1999; Gower et al. 2001; Hwang and Lee 2003; Landmann et al. 2004; Teng et al. 2004; Wang and Chamberlin 2004; Zhao et al. 2005). Progress is limited because of the complexity of the analysis. It involves dissection of all of the sequences around the gene of interest, which could be >10 kb long, to search for functional sequences. To facilitate the study of tissue-specific gene regulation in *C. elegans*, we use *C. elegans* muscle-specific gene expression as an example to explore the feasibility of identifying tissue-specific regulatory sequences through a computational approach. In *C. elegans*, muscle development has been an extensive area of research for a long time. Transcription factors of the basic helix-loop-helix class (*hlh-1*, *Ce-Twist*), the NK-2 class (*ceh-22*), and the T-box family (*tbx-2*, *mls-1*) have been shown to be critical for muscle specification and development (Okkema et al. 1993; Chen et al. 1994; Okkema and Fire 1994; Harfe and Fire 1998; Kostas and Fire 2002; Smith and Mango 2007). The promoter regions of several muscle-specific genes (*myo-1, myo-2, myo-3*, *unc-54*, *hlh-1*, and *ace-1*) have been studied in detail to identify important DNA regulatory sequences using sequence deletions or mutations (Okkema et al. 1993; Chen et al. 1994; Culetto et al. 1999). However, no general rules about the transcriptional regulatory mechanisms that control gene expression in muscle tissue have been identified.

Studies from various organisms have revealed a common theme that transcription factor binding sites tend to be interconnected and function together to confer a particular context-specific expression on the target gene. Those clusters of transcription factor binding sites form a regulatory module that can be located in the upstream, downstream, or intronic sequences and can be moved from their native context and still recapitulate a portion of the native expression pattern independent of their position and orientation to the basal promoter (Arnone and Davidson 1997). Modules have been shown to be very useful in studying temporal and spatial gene expression regulation. Modular structure of regulatory elements is widely present in higher eukaryotes (Kirchhamer et al. 1996; Arnone and Davidson 1997) and has been noted in *C. elegans* (Jantsch-Plunger and Fire 1994).

[1]Corresponding author.
E-mail stormo@genetics.wustl.edu; fax (314) 362-7855.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5989907.

17:348–357 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

Due to the time-consuming and labor-intensive nature of experimental approaches, many computational tools have been developed recently to facilitate the identification of regulatory modules. However, the predictive value of most of the methods is either unknown or less than satisfactory.

Here we describe a de novo computational method for accurate identification of regulatory sequences that confer muscle-specific gene expression, as well as experimental tests of the predicted modules. Comparisons of the predicted modules with experimentally characterized modules show high sensitivity and positive predictive value (PPV, defined as True Positives/All Predictions). A totals of 88% (22/25) of experimentally characterized modules are predicted, and 65% (30/46) of our predicted modules are located within experimentally defined regions. The rest of the predicted modules have not been tested for function, so the PPV could be much higher; it is already much higher than currently available algorithms. We developed a scoring system to predict the muscle specificity for any segment of DNA sequence. When applied to the whole genome, this method can help discriminate muscle genes from non-muscle genes. Because no information about known modules was used for the predictions, we expected the new predictions to have the same sensitivity and PPV. To examine this, we experimentally tested the functionality of 12 predicted modules. Of these 12 modules, three are located within known muscle gene promoters and nine are located in the promoters of genes with unknown expression patterns and unknown functions. Ten of the 12 tested modules drive gene expression in muscle tissue, demonstrating that our method is a valuable tool for guiding experimental design. Although we focus on muscle-specific gene expression in this work, we expect the method to be generally applicable to many other context-specific module identification tasks, because our method requires no prior knowledge other than a set of likely coexpressed orthologous genes. *C. elegans* muscle-specific module prediction tool can be accessed at http://ural.wustl.edu/software.html.

Results

Identification of regulatory motifs

Promoters are commonly defined as the DNA regions located upstream of the transcription start sites that contain the necessary binding elements for proper transcriptional regulation. In *C. elegans*, 60% of predicted intergenic regions will be fully included within a 2-kb upstream segment (Dupuy et al. 2004). The level of similarity between *C. elegans* and its relative *Caenorhabditis briggsae* decreases dramatically 1500 bp upstream of the predicted ATG for most genes with a long intergenic region (Dupuy et al. 2004). Even though some regulatory elements can be located in introns and/or 3′UTRs of genes (Okkema et al. 1993; Jantsch-Plunger and Fire 1994), including those regions in our study, could make computational identification of DNA motifs more difficult, because noise increases with increasing sequence length (Buhler and Tompa 2002; Wang and Stormo 2003). Therefore, we chose to focus on the upstream −2000 to −1 regions. We have used the translation start site (the 0 position) to select the candidate promoter regions because transcriptional start sites have not been determined for most *C. elegans* genes.

We used the program PhyloCon (Wang and Stormo 2003) for motif identification because comparisons suggest that it outperforms several previous motif-finding programs (Wang and Stormo 2003; MacIsaac et al. 2006). PhyloCon uses position weight matrix-based models (Stormo 2000) to represent ungapped DNA sequence motifs, and conserved motifs identified by this program represent potential regulatory elements. We collected a total of 122 *C. elegans* genes that are preferentially expressed in muscle tissue (Supplemental Table 1; details given in Methods section), 78 of which have defined *C. briggsae* orthologs (Supplemental Table 2). PhyloCon was run on the 2-kb upstream sequences of the 78 pairs of muscle genes to predict regulatory motifs, and a total of 18 unique motifs were identified (Table 1).

Muscle specificity of identified motifs

To identify motifs that are enriched in muscle gene promoters we calculated the Over Representation Index (ORI) (Bajic et al. 2004) for each motif (see Methods) using the rest of the genome as a background gene set. ORI takes into account not only the number of patterns found in sequences, but also the proportion of sequences in which the pattern is found. It reflects how much more probable it is to find a particular motif in the muscle-specific promoter set than in the background set. We define motifs that have an ORI >1.2 as muscle-specific motifs, and they are used later in module score calculations. From our catalog of 18 motifs, eight are designated as muscle-specific.

The top four motifs, ranked by ORI (Table 1), are similar to previously identified muscle-specific regulatory motifs (GuhaThakurta et al. 2002, 2004; Ao et al. 2004). Motif 1 (CTCTCTCTCTC) has almost the same consensus sequence as the binding site of transcription-factor TFII-I (currently known as GTF2I) in vertebrates, which binds to 5′-CTCACTCTCT-3′ (Clark et al. 1998). TFII-I family proteins play an important role in regulating muscle gene expression in humans (Polly et al. 2003). However, no *C. elegans* homolog was identified by BLAST. Motif 3 (CGCCRCCGCCKCC) is similar to the binding site of *Drosophila melanogaster* transcription factor Adf-1 (CCGCYGCYG YNGCCGV) in the TRANSFAC database (Matys et al. 2003). Homology search identified three genes in *C. elegans* that have sig-

Table 1. Predicted motifs ranked by over representation index (ORI)

Name	ORI	Consensus sequence	Note
1	3.60	CTCTCTCTCTC	Motif 2 (GuhaThakurta et al. 2002. 2004)
2	2.74	CTTCTTCTTCTTC	Motif 3 (GuhaThakurta et al. 2002, 2004)
4	2.21	RCACACAC	Daf-12 (Ao et al. 2004)
3	1.92	CGCCRCCGCCKCC	Motif 1 (GuhaThakurta et al. 2002, 2004)
16	1.47	CACTTCT	
18	1.27	CAATCRACAC	
17	1.24	TNGATCCATC	
15	1.23	ATGCCCT	
5	1.07	GCAAANAARGC	
6	1.06	WCTTTGM	
14	1.03	CTGACCG	
12	0.96	CCMAAAMC	
13	0.96	TCTGGTT	
9	0.95	CGTTTCG	
11	0.95	SACGTGG	
8	0.94	ACTGCAG	
7	0.92	YCAWTTTTC	
10	0.79	TTCCAGA	

ORI is calculated as described in the Methods. It reflects how more probable it is to find a motif in the muscle-specific promoter set than in the background set. The higher the ORI, the more enriched the motif is in muscle gene promoters.

nificant similarity to and belong in the same conserved orthologous groups (COG) as Adf-1. All of them have a MADF domain that directs sequence-specific DNA binding. Motif 6 (WCTTTGM) matches several similar matrices that belong to TCF/LEF family transcription factors that are a subfamily of HMG domain proteins that bind to WWCAAWG consensus sequences. It occurs at a similar level in the muscle gene promoters as in the background gene promoters. Therefore, our motif identification step recovered both known muscle-specific motifs as well as binding sites for common transcription factors.

Identification of muscle-specific regulatory modules in *C. elegans* promoter sequences

Currently, we do not have a good understanding on how motifs are organized to form modules. Modules may vary in the type of motifs, in the total number and the order of binding sites for each type of motif they contain. However, modules usually contain clusters of motifs, and this property has been used in various algorithms to identify regulatory modules (Wagner 1999; Berman et al. 2002; Markstein et al. 2002). In this study, we developed and tested a simple algorithm that is based on motif clustering and takes into account the general properties of well-studied regulatory modules in higher organisms. First, from many cases of well-studied regulatory modules in various organisms, regulatory modules usually consist of two to eight different regulatory motifs (Arnone and Davidson 1997). Therefore, we require that a regulatory module have at least two different motifs. Secondly, Wasserman and Fickett (1998) collected 18 well-characterized regulatory modules from human muscle genes. Most of the modules have at least two muscle-specific motif sites, which can be the sites of the same motif or of different motifs. Based on this information, we require that a regulatory module have at least two muscle-specific motif sites in order to be a muscle-specific regulatory module. Third, we require the distance between any two adjacent sites within a cluster to be ≤40 bp. Although this choice is somewhat arbitrary, the results are fairly insensitive to several reasonable choices of spacing between motifs (see Discussion). In summary, our definition of a muscle-specific regulatory module is a fragment of sequence that consists of clusters of motifs with intersite spaces ≤40 bp, and in which there are at least two different motifs and at least two muscle-specific binding sites (for details of the algorithm, see Methods).

Because some genes have alternative promoters, there are 138 different muscle gene promoters for the 122 muscle-specific genes. We applied this method on the 138 muscle gene promoters and identified 373 modules, an average of 2.7 modules per gene. The size of the modules ranges from 28 to 516 bp with a mean of 144 bp. Kirchhamer et al. (1996) collected 68 experimentally defined modules from *Drosophila* and mouse. Their size ranges from 40 bp to 8 kb, but they noted that the listed size was the length of DNA fragments used in gene transfer experiments and the actual size of the modules could be much smaller. The number of motifs in our predicted modules ranges from two to 12 with a mean of six. Well-studied modules have two to eight motifs with a mean of five (Arnone and Davidson 1997). Thus, our predicted modules share some general features with those well-studied modules.

Verification of regulatory modules

To evaluate the accuracy of the predicted modules we identified a total of 27 experimentally characterized modules in 16 gene promoters (Table 2). Of those 27 modules, one is located >2 kb upstream of the translation start site, outside the range of our predictions. Two of the modules overlap by >70% of their length (−370 to −686, −458 to −764 in gene *T18D3.4*) and it has not been tested whether the minimal overlapping region is sufficient for functionality, so they are treated as one module (−370 to −764) when calculating sensitivity and PPV. Therefore, there are a total of 25 experimentally characterized modules located in the regions we studied.

A comparison of our predicted modules to those experimentally characterized modules shows that they match closely. For example, *T18D3.4* encodes Myo-2, a pharyngeal-specific myosin heavy chain. The −17 to −239 region is defined as the minimal promoter that can drive reporter gene expression in pharyngeal muscles, while two overlapping 0.3-kb fragments (−370 to −686 and −458 to −764) are sufficient for pharyngeal muscle-specific enhancer activity (Okkema et al. 1993). We predicted three modules in the *T18D3.4* 2-kb upstream sequences that are located at −60 to −263, −430 to −515, and −562 to −733 upstream of the ATG start codon. Therefore, all three predicted modules are located within the experimentally defined regions (Fig. 1). In summary, for the 25 experimentally defined modules, our method correctly detected 88% (22/25). The definition of correct prediction is that the predicted modules overlap at least 50% with reported modules. In those 16 genes, our method predicted a total of 46 modules, of which 30 overlap with experimentally verified modules. Only one is located within a region shown not to be functional in muscle expression. Because the rest have not been tested, we cannot calculate the specificity of the prediction. However, the PPV of the prediction is at least 65% and could be as high as 98% if the rest of the predicted modules are all true positives. Supplemental Figure 2 shows the location of predicted and experimentally characterized modules for the entire set of genes. Also worth noting is that there are 15 experimentally defined modules with a length ≤ 500 bp. The distance from the ends of predicted modules to the ends of experimentally characterized modules ranges from 5 to 182 bp, and the average is 69 bp. These results demonstrate that the predicted regulatory modules are highly correlated with experimentally determined enhancers that direct gene expression in muscle tissues.

We performed simulations to estimate the statistical significance of obtaining the same sensitivity and PPV, given the promoter sequences and the known regulatory modules. We simulate the distribution of predicted modules in the promoters by randomly picking a start position for each module. The length and number of modules in each gene is kept the same as the predicted modules in that gene. The simulation is repeated 100,000 times and the sensitivity and PPV are calculated for each one. The average sensitivity is 48.8% with standard deviation of 7.8. The average PPV is 35.5% with standard deviation of 5.5. Therefore, the *P*-values of getting 88% sensitivity and 65% PPV are both much less than 0.001.

Detection of muscle genes on a genome scale

Another test of the accuracy of our module definitions is to use them to predict additional muscle genes. We developed a scoring system to measure the muscle specificity for each module using only the muscle-specific motif sites (see Methods). We expect that the higher the score, the more likely it is to be a muscle-specific module. By ranking all promoters by their scores we should be able to enrich for muscle genes. One difficulty of this

Table 2. Performance on muscle gene promoters

		Known module		Predicted module		
Gene ID	Name	Start	End	Start	End	Reference
B0304.1	*hlh-1*	−725	−949	−690	−831	Krause et al. 1994
		−1579	−1932	−1525	−1750	
		−457	−536	NP		
C02B8.4	*hlh-8*	−1	−315	−152	−339	Harfe and Fire 1998
				−1232	−1298	
C09D1.1a	*unc-89*	−1	−588	−66	−209	GuhaThakurta et al. 2004)
				−1210	−1258	
C36E6.5	*mlc-2*	−1	−400	−122	−195	GuhaThakurta et al. 2004
				−1895	−1986	
F07A5.7	*unc-15*	−1	−500	−39	−166	GuhaThakurta et al. 2004
				−958	−1054	
				−1492	−1523	
F11C3.3	*unc-54*	−61	−241	−66	−226	Okkema et al. 1993
				−1637	−1682	
				−1711	−1804	
F29F11.5	*ceh-22*	−18	−801	−36	−110	Vilimas et al. 2004
		−1436	−1554	−1309	−1544	
		−1583	−1808	−1607	−1644	
		−1794	−1922	NP		
F40E10.3	*csq-1*			−81	−207	Cho et al. 1999
		−260	−520	−371	−460	
F55B12.1	*ceh-24*	−1783	−1910	−1835	−1994	Harfe and Fire 1998
		−1989	−2443	Out of range		
F58A3.2a	*egl-15*	−1	−701	−237	−291	Harfe and Fire 1998
				−1381	−1589	
K12F2.1	*myo-3*	−328	−749	−297	−630	Okkema et al. 1993
		−1010	−1948	−1365	−1448	
R06C7.10	*myo-1*	−123	−500	−49	−350	Okkema et al. 1993
		−646	−1725	−1359	−1505	
				−1564	−1709	
T18D3.4	*myo-2*	−17	−239	−60	−263	Okkema et al. 1993
		−370	−686	−430	−515	
W09B12.1	*ace-1*	−564	−704	−613	−684	Culetto et al. 1999
		−1731	−2140	−1622	−1777NCP	
Y105E8B.1a	*tmy-1*	−1	−818	−10	−96	Kagawa et al. 1995
				−1541	−1600	
Y105E8B.1c	*tmy-1*	−1	−853	−74	−202	Anyanful et al. 2001
				−1347	−1506	
				−1540	−1669	

(NP) Not predicted. (NCP) Not correctly predicted.

assessment is that the expression pattern for most *C. elegans* genes is unknown. WormBase contains information about the tissue-expression pattern of 2576 genes. There are undoubtedly some omissions in these annotations, where some genes are expressed in tissues beside those listed, but it is likely to be largely correct and is the best data available for this assessment. For these 2576 genes, 1562 are either ubiquitously expressed or expressed in tissues other than muscle. We use these 1562 genes as the negative set. The set of well-characterized muscle genes that were not included in the training set, because we could not identify orthologs in *C. briggsae*, were used as a test set. We present the results using a Receiver Operator Characteristic (ROC) curve (Fig. 2) (Gribskov and Robinson 1996). For every possible choice of cutoff score, the *Y*-axis shows the fraction of true positives (known muscle genes) exceeding the cutoff, and the *X*-axis shows the fraction of false positives (known non-muscle-specific genes) exceeding the threshold. Any form of random predictions would result in points along the diagonal. The result demonstrates that our prediction is well above random, especially for the highest-scoring subset. For example, 30% of the muscle genes are detected at a threshold exceeded by <5% of the negative set, and 50% of the muscle genes are detected at a threshold exceeded by only 12% of the negative set. We use the area under the ROC curve (AUC) to measure how significant our prediction accuracy is. The AUC derived from the ROC curve is 0.7506, which is significantly different from 0.5 with $P << 0.0001$, indicating that our predictions are highly significant. However, there remain some well-characterized muscle genes that are not well predicted; their scores are not higher than the majority of the negative set genes. This means that we do not yet have a complete model that allows us to predict all muscle-specific gene expression. For some of these genes there is evidence that the muscle module occurs outside of the 2-kb promoter region we have used for scoring, but for others, we have to assume we are still missing some important features.

Will prior information help?

Our module predictions did not rely on any knowledge about experimentally defined modules, such as which genes contained them, where they were located, or which motifs they contained. We next examined whether the use of prior information about experimentally defined modules can identify a reduced set of motifs that is indispensable for module identification and can improve predictive performance.

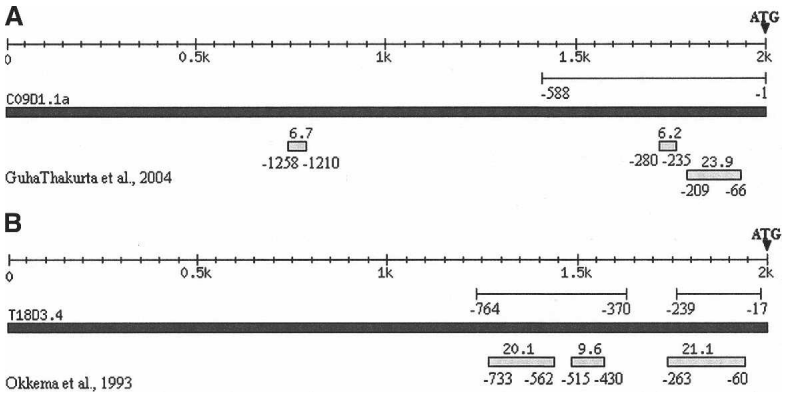

Figure 1. Two examples of comparison between predicted modules and experimentally defined modules. The lines *below* the scale represent the experimentally defined modules with start and end positions labeled. The black filled box represents the DNA sequence of corresponding gene. The end at the *right* side is −1 position of the gene. The small filled boxes *below* represent predicted modules with start and end positions labeled *below*. Black filled triangles indicate translational start sites. (*A*) For *C09D1.1a*, −1 to −588 bp upstream of ATG is an experimentally defined module. Our method predicted three modules and two are located within the first 588 bp (−66 to −209, −235 to −280). (*B*) *T18D3.4* has two experimentally defined modules, −17 to −239 and −370 to −764. We predicted three modules, and all three are located within experimentally defined regions.

First, we tested the performance of module prediction using only muscle-specific motifs. We first noticed that the sensitivity is greatly reduced compared with the prediction made with the full set of motifs. Varying the distance parameter from 20 to 100 bp, the sensitivity ranges from 52% to 72%, while using the full set of motifs has a sensitivity range from 80% to 96%. Secondly, the PPV (from 61.8% to 74.3%) is comparable to the prediction made with the full set of motifs (from 60.5% to 77.4%). Using this motif set to perform genomic predictions does not improve the performance, as determined by the ROC curve of the 44 test set muscle-specific genes (Supplemental Fig. 3). This suggests that some of the non-muscle-specific motifs are important components of muscle-specific modules. We next performed experiments to find a subset of motifs to regain the prediction sensitivity with the same or higher level of PPV. By adding back combinations of one, two, or three non-muscle-specific motifs and using various distance parameters ranging from 20 to 100 bp, we find that there are six cases in which we can obtain both higher sensitivity and higher PPV (Supplemental Table 3). In all cases, motif 6 (WCTTTGM) is included in the motif set. We used three motif sets that give the highest sensitivity and PPV to perform genomic prediction, and plotted the ROC curve of the 44 test set muscle genes. The results suggest that the predictive performances are all comparable to, or worse than, the original set of motifs (Supplemental Fig. 3). Therefore, training on known modules can improve the performance on the training set, but this must be due to overfitting, because it does not improve the genomic predictions in any significant way. These results demonstrate that (1) our method for module identification does not need prior information in order to make high quality predictions; (2) our method is robust; (3) the initial step of motif prediction and redundant motif elimination effectively identifies motifs that are important for regulating muscle-specific gene expression.

Experimental verification of predicted modules

All of the statistical analyses suggest that our method generated high-quality predictions. To test the predictive value of the method on unknown modules and the usefulness in guiding experimental designs, we performed four different types of experiments.

First, we tested our predictive powers by locating the regulatory regions of three genes that are known to be muscle-specific genes, but whose promoters have not been subjected to comprehensive functional analyses. Our results confirmed that our predictions are correct in all three cases. *C02D4.2* (*ser-2*) has at least three alternative promoters that drive *C02D4.2* expression in a set of neurons, as well as pharyngeal cells and head muscles (Tsalik et al. 2003). We predicted three modules in *C02D4.2a* 2-kb upstream region (−91 to −382, −1557 to −1716, and −1769 to −1882). We verified the function of the first predicted module by determining that the first 512 bp upstream of the ATG is sufficient to drive *gfp* expression only in the head muscle cells (data not shown). Similarly, DNA sequences encompassing the first predicted modules of *C33G3.1a* (*dyc-1*) and *F08B6.2* (*gpc-2*) both drive reporter gene expression in the corresponding muscle cells (Table 3; data not shown).

Second, we tested whether our predictions help to identify muscle-expressing genes in the genome. We randomly picked eight genes of unknown function and unknown expression pattern from the top-ranking predicted muscle genes (ranked from 1 to 198 in the genomic ranking, Table 3). For each gene we assayed whether the minimal upstream sequences encompassing the first predicted modules could drive gene expression in the muscle tissue. Table 3 shows the list of genes tested, as well as the

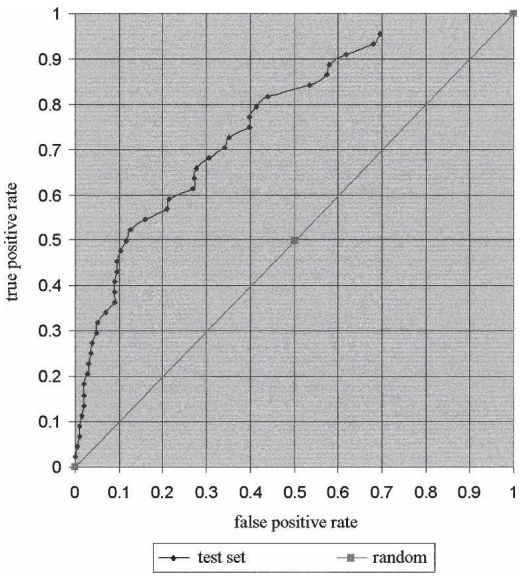

Figure 2. ROC curves of muscle gene prediction. Genomic genes are ranked by their muscle-specificity score. We plotted the ROC curves of the set of well-characterized muscle-specific genes that were not used for motif identification (44 test set). The diagonal line represents the result of random guessing. The *Y*-axis is the fraction of true positives exceeding the cutoff for every cutoff value. The *X*-axis is the fraction of true negatives that exceed the same cutoff.

Table 3. Experimental validation of predicted modules

	Gene ID	Genomic rank	Length of IGS	Predicted module	Sequence tested	Gene expression pattern
Known gene	*C02D4.2a*	141	~17 kb	−91, −382	−1, −512	Head muscle only
	C33G3.1a	372	>10 kb	−139, −437	−1, −520	Body wall and vulva muscles
	F08B6.2.1	2216	~1.3 kb	−329, −647	−1, −770	Muscle cells and neurons
Unknown Gene	*C01B7.1b*	1	~2.6 kb	−6, −524	−1, −667	Pharyngeal muscle, neuron
	C01B7.3	2	~2.6 kb	−85, −417	−1, −553	No green
	C10G11.7	7	~9.5 kb	−4, −625	−1, −741	Exclusively in neurons
	F45D3.2	21	>9 Kb	−153, −541	−1, −697	Body wall muscle
	W06H8.6	23	>7 Kb	−256, −591	−1, −675	Body wall, pharyngeal and vulva muscle, H cell
	T28B8.1.1	114	~4 kb	−56, −541	−1, −597	Head muscle, pharyngeal muscle, vulva muscle, neurons
	K10G6.3	173	>10 kb	−378, −847		
	Long form				−1, −929	Pharyngeal muscle, neuron
	Short form				−1, −378	NO
	F27D4.2	198	~11 kb	−491, −1041		
	Long form				−1, −1212	Body wall and pharyngeal muscle, intestine
	Short form				−1, −467	NO
	W06H8.6	23	>7 Kb	−746, −1183	−675, −1333	Pharyngeal muscle

genomic rank of the genes, the location of the predicted modules, and the observed expression patterns. *C01B7.3* and *C01B7.1* share the 2.6-kb intergenic sequences. *C01B7.3* is a predicted gene with no RNAi phenotype and no hit in a BLASTP search in the genome of *C. briggsae*, *Caenorhabditis remanei*, *Anopheles gambiae*, *D. melanogaster*, *Rattus norvegicus*, *Homo sapiens*, *C. elegans*, and *Saccharomyces cerevisiae* (WormBase http://www.wormbase.org/). In our experiment, the 553-bp *C01B7.3* promoter did not give any expression pattern. Therefore, *C01B7.3* is likely to be a falsely annotated gene. For the remaining seven genes, six are muscle genes, while the minimal promoter region of *C10G11.7* drives reporter gene expression exclusively in the neurons (Fig. 3A–L). It is known that muscle genes and neuronal genes share some regulatory elements (Wasserman and Fickett 1998) and 45% of our muscle-specific genes are also expressed in neurons. If we include neuronal genes as positive, 87.5% of our genes are correctly predicted.

Third, we tested the functionality of modules located further upstream by deletion analysis. The first two predicted modules in *K10G6.3* are clustered at −378 to −847. A DNA fragment containing this region drives *gfp* expression mainly in neurons and occasionally in the pharyngeal muscles (Fig. 3N). Deletion of this region results in complete loss of *gfp* expression. The first predicted module in *F27D4.2* is located at −491 to −1041. A DNA fragment including the predicted module drives reporter gene expression in the pharyngeal muscle (Fig. 3C), body wall muscle (Fig. 3D), and intestine, whereas deletion of the predicted module from the DNA results in loss of *gfp* expression.

Fourth, we tested the enhancer activity of a predicted module. *W06H8.6* is a gene with unknown function and unknown expression pattern that has an upstream sequence >7 kb. In the *W06H8.6* 2-kb promoter sequence, six modules were predicted. The first one is located at −256 to −591 and the first 675 bp upstream of ATG drives reporter gene expression in body wall muscle (Fig. 3J), vulva muscle (Fig. 3K), and pharyngeal muscle (Fig. 3L), as reported above. Another three are located between −764 and −1183 with intermodule distance of around 40 bp. We tested the functionality of this cluster of modules by introducing the DNA fragment upstream of a minimal *pes-10* promoter (Fire et al. 1990) and examined its ability to activate reporter gene expression. The tested DNA fragment drives reporter gene expression only in the pharyngeal muscle cells (Fig. 3M).

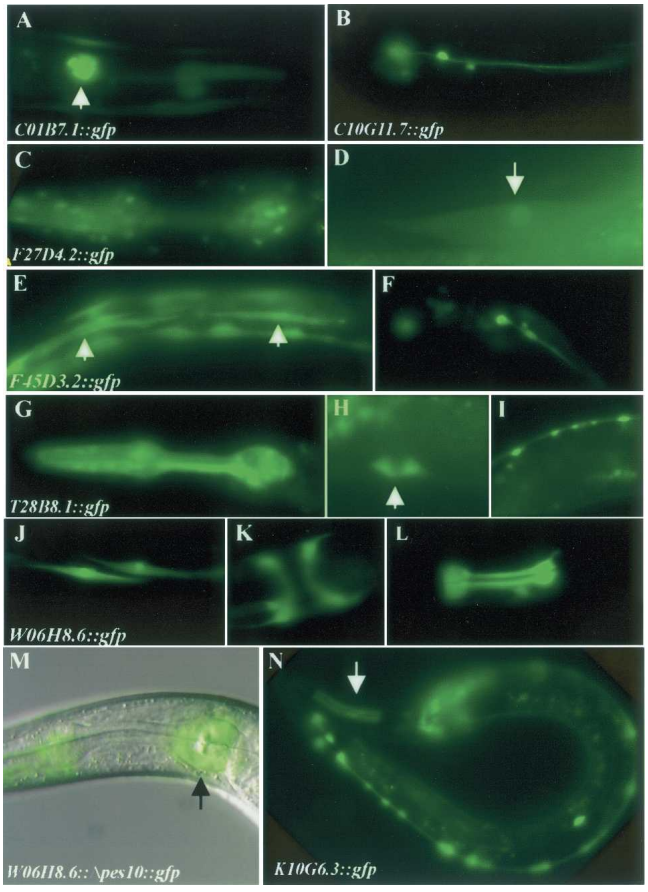

Figure 3. GFP expression pattern driven by DNA sequences encompassing the predicted modules. (*A*) *C01B7.1*::*gfp* expression in the pharyngeal muscle. (*B*) *C10G11.7*::*gfp* expression in neurons. *F27D4.2*::*gfp* expression in pharyngeal muscle (*C*) and body wall muscles (*D*, arrow). *F45D3.2*::*gfp* expression in the body wall muscle (*E*, arrows) and neurons (*F*). *T28B8.1*::*gfp* expression in the pharyngeal muscle (*G*), vulva muscle (*H*, arrow), and neurons (*I*). *W06H8.6*::*gfp* expression in the body wall muscle (*J*), vulva muscle (*K*), pharyngeal muscle (*L*), and H cells (data not shown). (*M*) A merge of DIC image and fluorescent image taken for the same worm showing *W06H8.6*::Δ*pes10*::*gfp* expression in the pharyngeal muscle cells. (*N*) *K10G6.3*::*gfp* expression in the pharyngeal muscle (arrow) and neurons.

Although both the promoter and the more upstream module of *W06H8.6* directed gene expression in muscle cells, each is expressed in a unique subset of muscle cells.

In summary, we tested the functionality of 12 predicted modules. Ten of them drive gene expression in muscle tissues and one of them is involved in gene expression in neuronal cells. The remaining one showed no expression and may not even correspond to a true gene. This gives a positive predictive value of 83%, and 92%, if we count neuronal regulatory modules as positive. Generally, it takes many similar experiments to dissect the long promoter sequences to identify the functional sequences of a single gene. For the genes we tested, several of them have very long upstream sequences. For example, the upstream sequence of *F45D3.2*, *W06H8.6*, and *F27D4.2* are 9, 7, and 11 kb, respectively. These results demonstrate that our method is able to both predict unknown genes that are expressed in muscle cells and to reduce the important functional domains, which contain the essential modules, to much smaller regions.

Discussion

The accurate identification of regulatory modules within a genomic sequence would be very useful for the study of gene regulation. However, identifying modules experimentally is a time-consuming and labor-intensive process. We developed a computational approach to predict muscle-specific *cis*-regulatory modules in *C. elegans* and performed experimental evaluations of their accuracy. Analysis of the in vivo activity of 12 predicted modules, of which 10 showed the predicted activity, demonstrates the utility of our approach.

We chose muscle genes for this study because muscle has been a fertile ground for molecular genetics studies with *C. elegans* for three decades. Most of the work focused on the organization, structure, and function of muscle fibers and muscle cells (Moerman and Fire 1997; Moerman and Williams 2006). Recent work identified two genes that are involved in muscle cell fate specification (Kostas and Fire 2002; Smith and Mango 2007). However, the molecular mechanisms that control muscle cell fate specification and differentiation remain unclear. Here we demonstrate a computational approach that can identify motifs and their combinations into regulatory modules, which is very useful in identifying muscle-expressing genes. We tested eight genes of unknown expression pattern and unknown function, which we predicted to have modules for muscle expression. Six of those modules did, indeed, cause expression in muscle cells, while one drove expression in neurons and another showed no expression pattern. In total, we tested 12 predicted modules, with 10 showing activity in regulating muscle gene expression, which gives a PPV of 83%. Many of those were in segments directly upstream of the gene, consistent with *C. elegans* regulatory regions being compact. But in two cases, K10G6.3 and F27D4.2, we showed that the immediate upstream region was not sufficient for muscle expression, but that inclusion of a predicted module further upstream was. In another case, W06H8.6, we showed that two predicted modules, one immediately upstream and another more distant one, were each sufficient to drive muscle expression, but with different expression patterns.

Although this study focused on modules for muscle expression, we did not use any muscle-specific characteristics, and we expect that our method would work equally well for other tissue-specific expression patterns. The approach is quite simple and requires very little prior information, including no initial information about motifs. The input is merely a set of *C. elegans* genes known to share a particular expression pattern and their orthologs in another *Caenorhabditis* genome, so that the program PhyloCon could identify significant motifs. We then used the promoters of non-muscle genes to identify which motifs were muscle specific and which were general. The set of motifs were then combined into predicted modules based on characteristics of a few well-characterized modules found in human, mouse, rat, fly, and sea urchin (Arnone and Davidson 1997; Wasserman and Fickett 1998), namely, that there should be at least two different motifs within the module and at least two occurrences of muscle-specific motifs. The one parameter we explored was the spacing between motifs within a module, but we found the results to be quite consistent over ranges from 20 to 75 bp (Supplemental Table 4); longer spacing often predicted entire upstream regions to be a single module, which is not very useful. We do not specify a particular window size for a module, and they can vary considerably in length. We also do not specify a minimum score, although the score, which is based only on the content of the muscle-specific motifs, is useful for ranking the predicted modules, and the results show that the highest-scoring promoters are the most enriched in muscle-specific genes (Fig. 2). In fact, the score appears to reflect the strength of the module in driving muscle gene expression based on the few experimentally determined modules with quantitative comparisons of activity. For example, Okkema et al. (1993) identified two modules in *myo-1* gene (*R06C7.10*): a strong proximal enhancer located at −123 to −500 and a weak distal enhancer located at −646 to −1752. We predicted three modules. The highest-scoring module (score 42.1) is located near the start site, corresponding to the proximal enhancer, and the two lower-scoring modules (score 18.9 and 11.7) are located distally, corresponding to the weaker enhancer (Supplemental Fig. 2). In *F29F11.5*, three modules located between −1436 and −1922 upstream of ATG were characterized (DE1, DE2, and DE3) with DE3 showing the strongest enhancer activity (Vilimas et al. 2004). Modules were correctly predicted for DE2 and DE3, with the latter having a higher score (Supplemental Fig. 2). Low-score modules could be functional modules as exampled in Supplemental Figure 2.

While these results demonstrate the utility of our approach, we are still far from having a precise and completely accurate predictor of muscle expression patterns. Two of the 12 predicted modules we tested were not correct. From the ROC curve (Fig. 2) it can be seen that high-scoring promoters are highly enriched in muscle-specific genes, but there are a few non-muscle genes that also have high scores, and there are several muscle-specific genes with only low-scoring modules that do not distinguish well from non-muscle genes. Furthermore, we have only attempted to predict muscle expression in general, rather than for specific classes of muscles. Among the tested modules we see several distinct patterns that include specific subsets of muscles from the head, body wall, vulva, and pharynx, as well as some that also cause expression in subsets of neurons. More work is needed before we can fully model more specific expression patterns. For example, in this study, we have not considered possible modules occurring within introns or downstream of the genes, even though we know of such examples (Jantsch-Plunger and Fire 1994). Nor have we considered the phenomenon that clusters of nearby genes may all be activated coordinately, perhaps through the modification of local chromatin domains (Roy et al. 2002). The recent release of the *C. remanei* genome sequence (ftp://

ftp.wormbase.org/pub/wormbase/genomes/remanei/) will increase our power to detect conserved regulatory motifs using methods such as PhyloCon and PhyloNet (Wang and Stormo 2003, 2005). In addition, comprehensive analysis is now ongoing to determine the complete repertoire of *C. elegans* TFs and their binding-site motifs (Reece-Hoyes et al. 2005). Together, we expect that these additional data will allow for more comprehensive characterization of regulatory interactions and aid in the determination of the complete regulatory network of a model metazoan.

Methods

Identification of *C. elegans* muscle genes and orthologs in *C. briggsae*

In this study, we define muscle-specific genes as those that are only expressed in the muscle tissue or expressed in at most two other tissues. We identified a total of 122 *C. elegans* muscle-specific genes from searching the WormBase (Chen et al. 2005) expression pattern database (http://www.wormbase.org/) and from previous work (GuhaThakurta et al. 2002). *C. briggsae* orthologs for 78 of the 122 genes were obtained from WormBase. The *C. elegans* and *C. briggsae* chromosomal sequence and the gene structures were downloaded from the WormBase ftp-site (ftp://ftp.wormbase.org/pub/wormbase/genomes/, WS123). These were then used to obtain −2000 to −1 upstream regions of muscle-specific genes, as well as an upstream region of all *C. elegans* genes (22,247).

Identification of putative regulatory motifs and elimination of redundant motifs

PhyloCon (Wang and Stormo 2003) program was run on the upstream sequences (−2000 to −1) of the 78 pairs of *C. elegans* and *C. briggsae* orthologous muscle genes. We took the best matrix from each run of PhyloCon, masked all of the incidences of the identified motif in the input file, and repeated until no additional significant motifs were identified. The experiments were performed using various parameters (Wang and Stormo 2003), and motifs identified in all experiments were pooled together. To determine whether any two-position weight matrices were similar, we tested whether two motifs overlap significantly in promoter sequences, as determined by a χ^2 test on simulated data. If two motifs overlap significantly, they were considered redundant motifs, and the one with lower information content was removed.

Calculation of over-representation index

Given a weight matrix, the Patser program calculates the probability of observing a sequence with a particular score or greater (Staden 1989; Hertz and Stormo 1999) and determines the default cutoff score based on that *P*-value. Therefore, a "site" corresponding to a particular motif (weight matrix) is a subsequence that is identified by the Patser program using the cutoff appropriate for each motif.

We adopted the concept of over-representation of a particular pattern in one group of sequences with regard to another group of sequences from Bajic et al. (2004). They define it as:

$$\mathrm{ORI(Mi)} = \frac{\mathrm{Density}_{\mathrm{specific}}(\mathrm{Mi})}{\mathrm{Density}_{\mathrm{nonspecific}}(\mathrm{Mi})} \times \frac{\mathrm{Proportion}_{\mathrm{specific}}(\mathrm{Mi})}{\mathrm{Proportion}_{\mathrm{nonspecific}}(\mathrm{Mi})} \tag{1.1}$$

where Mi is the i$^{\mathrm{th}}$ motif. $\mathrm{Density}_{\mathrm{specific}}(\mathrm{Mi})$ is the density at which this motif is found in muscle-specific promoter sequences, and $\mathrm{Density}_{\mathrm{nonspecific}}(\mathrm{Mi})$ is the density at which this motif is found in nonspecific sequences. Density is the number of sites of motif i in a sequence of unit length. $\mathrm{Proportion}_{\mathrm{specific}}$ is the proportion of muscle-specific promoters that has the motif i. $\mathrm{Proportion}_{\mathrm{nonspecific}}$ is the proportion of nonspecific promoters that has the motif i. This can be rewritten as:

$$\mathrm{ORI(Mi)} = \frac{\dfrac{\mathrm{NumSite}_{\mathrm{s}}(\mathrm{Mi})}{\mathrm{Total\ Length}_{\mathrm{specific}}} \times \dfrac{\mathrm{N}_{\mathrm{s}}(\mathrm{Mi})}{\mathrm{Total\ Promoter}_{\mathrm{specific}}}}{\dfrac{\mathrm{NumSite}_{\mathrm{ns}}(\mathrm{Mi})}{\mathrm{Total\ Length}_{\mathrm{nonspecific}}} \times \dfrac{\mathrm{N}_{\mathrm{ns}}(\mathrm{Mi})}{\mathrm{Total\ Promoter}_{\mathrm{nonspecific}}}} \tag{1.2}$$

$\mathrm{NumSite}_{\mathrm{s}}$ is the number of sites of motif i found in muscle-specific promoter sequences, while $\mathrm{NumSite}_{\mathrm{ns}}$ is the number of sites of motif i found in nonspecific sequences; $\mathrm{TotalLength}_{\mathrm{specific}}$ is the total length of muscle-specific promoter sequences and $\mathrm{TotalLength}_{\mathrm{nonspecific}}$ is the total length of nonspecific sequences; N_{s} is the number of muscle-specific promoter sequences where motif i is found, N_{ns} is the number of nonspecific sequences where motif i is found; $\mathrm{TotalPromoter}_{\mathrm{specific}}$ is the total number of muscle-specific promoter and $\mathrm{TotalPromoter}_{\mathrm{nonspecific}}$ is the total number of nonspecific sequences, respectively. We use all *C. elegans* genes other than the 138 muscle gene promoters as nonspecific background sequences.

Searching for *cis*-regulatory modules

To search for clusters of motifs, we first identify all of the sites for all of the motifs using Patser. Then, we scan the sequence from 5′ to the 3′ end starting from the first site in the sequence. If the next site is less than the cutoff distance away, it is considered to be in the same cluster as the first site. Then, the third site is considered and the distance between it and the second site is calculated. This processes continues until a site is encountered that is too far away from the previous site (exceeds the distance cutoff). This cluster of motifs is a putative regulatory module. Then, we check whether this cluster fits the criteria of muscle-specific module (having at least two types of motifs and two muscle-specific sites). If it fits, it is kept as a muscle-specific regulatory module.

Calculation of module score and promoter score

For a given DNA sequence, the combined probability–proportionality value of multiple motifs is calculated as described (GuhaThakurta et al. 2004). It measures the likelihood that each TF binds at least one of its binding sites in the given sequence. We apply this calculation on each predicted module rather than the whole sequence to calculate the combined probability–proportionality value for each module:

$$P^{\mathrm{module}} = \prod_{m=1}^{n} P_{\mathrm{m}}^{\mathrm{module}} \tag{1.3}$$

where *m* denotes all of the motifs that exist in the module and *n* is the total number of different motifs. $P_{\mathrm{m}}^{\mathrm{module}}$ is the probability–proportionality value for motif *m* in a given module *module* calculated as described (GuhaThakurta et al. 2004). This treatment is likely oversimplified given the known cooperative binding of transcription factors to promoter elements. However, this does not affect module prediction, it only affects the ranking of genes when we try to discriminate muscle genes from non-muscle genes, and this simplified approach has produced meaningful results. The score for a regulatory module is calculated as log of the combined probability.

$$S^{\text{module}} = \log P^{\text{module}} \quad (1.4)$$

If a promoter region has more than one identified regulatory module, the muscle-specificity score for the promoter is the sum of the score of all the modules it has

$$S^{\text{promoter}} = \sum_{i=1}^{n} S_i^{\text{module}} \quad (1.5)$$

where n is the total number of modules in the promoter.

Genome-wide searches

We retrieved 2 kb of upstream sequences from all of the genes in the *C. elegans* genome (22,247). A muscle-specificity score is calculated for each gene promoter as described above. The promoters were then ranked by the score. If a gene has multiple promoters, we take the highest score and ranking of that gene.

Construction of plasmids and GFP expression analysis

To test the predicted modules close to translational start codons, gene-specific primers were used to amplify the corresponding sequences from fosmid DNAs (Geneservice Ltd). PCR products were cloned into a promoterless GFP vector pLS43 (Guha-Thakurta et al. 2004) with nuclear localization signals. Transgenic *C. elegans* were made as described (Mello et al. 1991) using the collagen gene *rol-6* as a coinjection marker. Rolling GFP-expressing progeny were isolated and studied for in vivo GFP expression.

To test the enhancer activity of more distant predicted modules, PCR products were cloned into pPD107.94 (Δ*pes-10* minimal promoter, a gift from Andrew Fire, Stanford University School of Medicine) (Fire et al. 1990). The construct is used to make transgenic animals for GFP expression study.

Acknowledgments

We thank Ting Wang for assistance with the PhyloCon program and helpful discussions. We also thank Michael L. Nonet, Andrew Fire, and Susan E. Mango for providing reagents used in this work, and Dr. Frank E. Harrell Jr. for helping with statistical analysis of the predictions. This work was supported by NIH grants HG00249, and G.Z. was supported by NIH institutional training grant 5 T32 HG000045-08 and National Institute of General Medical Sciences NRSA service award 1 F32 GM73444-01.

References

Anyanful, A., Sakube, Y., Takuwa, K., and Kagawa, H. 2001. The third and fourth *tropomyosin* isoforms of *Caenorhabditis elegans* are expressed in the pharynx and intestines and are essential for development and morphology. *J. Mol. Biol.* **313:** 525–537.

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.

Arnone, M.I. and Davidson, E.H. 1997. The hardwiring of development: Organization and function of genomic regulatory systems. *Development* **124:** 1851–1864.

Bajic, V.B., Choudhary, V., and Hock, C.K. 2004. Content analysis of the core promoter region of human genes. *In Silico Biol.* **4:** 109–125.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B. 2002. Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci.* **99:** 757–762.

Buhler, J. and Tompa, M. 2002. Finding motifs using random projections. *J. Comput. Biol.* **9:** 225–242.

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.

Chen, N., Harris, T.W., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Canaran, P., Chan, J., Chen, C.K., et al. 2005. WormBase: A comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Res.* **33:** D383–D389.

Cho, J.H., Eom, S.H., and Ahnn, J. 1999. Analysis of *calsequestrin* gene expression using green fluorescent protein in *Caenorhabditis elegans*. *Mol. Cells* **9:** 230–234.

Clark, M.P., Chow, C.W., Rinaldo, J.E., and Chalkley, R. 1998. Multiple domains for initiator binding proteins TFII-I and YY-1 are present in the initiator and upstream regions of the rat XDH/XO TATA-less promoter. *Nucleic Acids Res.* **26:** 2813–2820.

Culetto, E., Combes, D., Fedon, Y., Roig, A., Toutant, J.P., and Arpagaus, M. 1999. Structure and promoter activity of the 5′ flanking region of *ace-1*, the gene encoding acetylcholinesterase of class A in *Caenorhabditis elegans*. *J. Mol. Biol.* **290:** 951–966.

Dupuy, D., Li, Q.R., Deplancke, B., Boxem, M., Hao, T., Lamesch, P., Sequerra, R., Bosak, S., Doucette-Stamm, L., Hope, I.A., et al. 2004. A first version of the *Caenorhabditis elegans* Promoterome. *Genome Res.* **14:** 2169–2175.

Fire, A., Harrison, S.W., and Dixon, D. 1990. A modular set of lacZ fusion vectors for studying gene expression in *Caenorhabditis elegans*. *Gene* **93:** 189–198.

Gilleard, J.S., Shafi, Y., Barry, J.D., and McGhee, J.D. 1999. ELT-3: A *Caenorhabditis elegans* GATA factor expressed in the embryonic epidermis during morphogenesis. *Dev. Biol.* **208:** 265–280.

Gower, N.J., Temple, G.R., Schein, J.E., Marra, M., Walker, D.S., and Baylis, H.A. 2001. Dissection of the promoter region of the inositol 1,4,5-trisphosphate receptor gene, *itr-1*, in *C. elegans*: A molecular basis for cell-specific expression of IP3R isoforms. *J. Mol. Biol.* **306:** 145–157.

Gribskov, M. and Robinson, N.L. 1996. The use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.* **20:** 25–34.

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.* **7:** 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.

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.

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

Hwang, S.B. and Lee, J. 2003. Neuron cell type-specific *SNAP-25* expression driven by multiple regulatory elements in the nematode *Caenorhabditis elegans*. *J. Mol. Biol.* **333:** 237–247.

Jantsch-Plunger, V. and Fire, A. 1994. Combinatorial structure of a body muscle-specific transcriptional enhancer in *Caenorhabditis elegans*. *J. Biol. Chem.* **269:** 27021–27028.

Kagawa, H., Sugimoto, K., Matsumoto, H., Inoue, T., Imadzu, H., Takuwa, K., and Sakube, Y. 1995. Genome structure, mapping and expression of the tropomyosin gene *tmy-1* of *Caenorhabditis elegans*. *J. Mol. Biol.* **251:** 603–613.

Kamachi, Y., Uchikawa, M., and Kondoh, H. 2000. Pairing SOX off: With partners in the regulation of embryonic development. *Trends Genet.* **16:** 182–187.

Kirchhamer, C.V., Yuh, C.H., and Davidson, E.H. 1996. Modular *cis*-regulatory organization of developmentally expressed genes: Two genes transcribed territorially in the sea urchin embryo, and additional examples. *Proc. Natl. Acad. Sci.* **93:** 9322–9328.

Kostas, S.A. and Fire, A. 2002. The T-box factor MLS-1 acts as a molecular switch during specification of nonstriated muscle in *C. elegans*. *Genes & Dev.* **16:** 257–269.

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.

Landmann, F., Quintin, S., and Labouesse, M. 2004. Multiple regulatory elements with spatially and temporally distinct activities control the expression of the epithelial differentiation gene *lin-26* in *C. elegans*. *Dev. Biol.* **265:** 478–490.

Li, R., Pei, H., and Watson, D.K. 2000. Regulation of Ets function by protein–protein interactions. *Oncogene* **19:** 6514–6523.

MacIsaac, K.D., Wang, T., Gordon, D.B., Gifford, D.K., Stormo, G.D., and Fraenkel, E. 2006. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* **7:** 113.

Markstein, M., Markstein, P., Markstein, V., and Levine, M.S. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl. Acad. Sci.* **99:** 763–768.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., et al. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31:** 374–378.

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

Moerman, D.G. and Fire, A. 1997. Muscle: Structure, function, and development. In C. elegans *II* (eds. D.L. Riddle et al.), pp. 147–184. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Moerman, D.G. and Williams, B.D. 2006. Sarcomere assembly in *C. elegans* muscle. In *WormBook* (ed. T.C.e.R. Community). WormBook.

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., 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.

Polly, P., Haddadi, L.M., Issa, L.L., Subramaniam, N., Palmer, S.J., Tay, E.S., and Hardeman, E.C. 2003. hMusTRD1alpha1 represses MEF2 activation of the *troponin I* slow enhancer. *J. Biol. Chem.* **278:** 36603–36610.

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.

Remenyi, A., Scholer, H.R., and Wilmanns, M. 2004. Combinatorial control of gene expression. *Nat. Struct. Mol. Biol.* **11:** 812–815.

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.

Smith, P.A. and Mango, S.E. 2007. Role of T-box gene *tbx-2* for anterior foregut muscle development in *C. elegans*. *Dev. Biol.* **302:** 25–39.

Spieth, J., MacMorris, M., Broverman, S., Greenspoon, S., and Blumenthal, T. 1988. Regulated expression of a vitellogenin fusion gene in transgenic nematodes. *Dev. Biol.* **130:** 285–293.

Staden, R. 1989. Methods for calculating the probabilities of finding patterns in sequences. *Comput. Appl. Biosci.* **5:** 89–96.

Stormo, G.D. 2000. DNA binding sites: Representation and discovery. *Bioinformatics* **16:** 16–23.

Teng, Y., Girard, L., Ferreira, H.B., Sternberg, P.W., and Emmons, S.W. 2004. Dissection of *cis*-regulatory elements in the *C. elegans* Hox gene *egl-5* promoter. *Dev. Biol.* **276:** 476–492.

Tsalik, E.L., Niacaris, T., Wenick, A.S., Pau, K., Avery, L., and Hobert, O. 2003. LIM homeobox gene-dependent expression of biogenic amine receptors in restricted regions of the *C. elegans* nervous system. *Dev. Biol.* **263:** 81–102.

Vilimas, T., Abraham, A., and Okkema, P.G. 2004. An early pharyngeal muscle enhancer from the *Caenorhabditis elegans* ceh-22 gene is targeted by the Forkhead factor PHA-4. *Dev. Biol.* **266:** 388–398.

Wagner, A. 1999. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* **15:** 776–784.

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, T. and Stormo, G.D. 2005. Identifying the conserved network of *cis*-regulatory sites of a eukaryotic genome. *Proc. Natl. Acad. Sci.* **102:** 17400–17405.

Wasserman, W.W. and Fickett, J.W. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278:** 167–181.

Zhao, Z., Fang, L., Chen, N., Johnsen, R.C., Stein, L., and Baillie, D.L. 2005. Distinct regulatory elements mediate similar expression patterns in the excretory cell of *Caenorhabditis elegans*. *J. Biol. Chem.* **280:** 38787–38794.

Received September 25, 2006; accepted in revised form December 12, 2006.