1-1-2011

Discovering Conserved cis-Regulatory Elements That Regulate Expression in Caenorhabditis elegans

Nnamdi Ihuegbu
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Computational and Systems Biology

Dissertation Examination Committee:
Tim Schedl, Chair
Jeremy Buhler
Joseph Corbo
Stephen Kornfeld
Gary Stormo
Ting Wang

DISCOVERING CONSERVED *cis*-REGULATORY ELEMENTS

THAT REGULATE EXPRESSION IN *Caenorhabditis elegans*

by

Nnamdi Ihuegbu

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2011

Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Discovering conserved *cis*-Regulatory elements that regulate expression in *Caenorhabditis elegans*

By

Nnamdi Ihuegbu

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2011

Professor Tim Schedl, Chairperson

The aim of this dissertation is two-fold: (1) To catalog all *cis*-regulatory elements within the intergenic and intronic regions surrounding every gene in *C.elegans* (i.e. the regulome) and (2) to determine which *cis*-regulatory elements are associated with expression under specific conditions. We initially use PhyloNet to predict conserved motifs with instances in about half of the protein-coding genes. This initial first step was valuable as it recovered some known elements and cis-regulatory modules. Yet the results had a lot of redundant motifs and sites, and the approach was not efficiently scalable to the entire regulome of *C. elegans* or other higher-order eukaryotes. Magma (Multiple Aligner of Genomic Multiple Alignments) overcomes these shortcomings by using efficient clustering and memory management algorithms. Additionally, it implements a fast greedy set-cover solution to significantly reduce redundant motifs.

These differences make Magma ~70 times faster than PhyloNet and Magma-based predictions occur near ~99% of all *C. elegans* protein-coding genes. Furthermore, we show tractable scaling for higher-order eukaryotes with larger regulomes. Finally, we demonstrate that a Magma-predicted motif, which represents the binding specificity for HLH-30, plays a critical role in the host-defense to pathogenic infections. This novel finding shows that *hlh-30(-)* animals are more susceptible to *S. aureus* and *P. aeruginosa* than their wild type counterparts.

Acknowledgments

There are many people who played crucial roles in my academic success during these past 6 years. Although this is not meant to be an exhaustive list, I will like to highlight a few of these people. I will like to thank Gary Stormo for being an excellent and supportive mentor. Although he is a very busy person, he has still made time to mentor me on a variety of subjects from algorithm optimization strategies to advice on life after graduate school. It has truly been a pleasure and an honor to be an apprentice of his. I will also like to thank current and former members of the Center for Genome Sciences and the Stormo lab for their friendship and lab meeting discussions which have enlightened me, challenged me, and ultimately made me a better critical thinker and scientist. In particular I will like to thank Aaron Spivak, Barrett Foat, Larry Schriefer, Lee Tessler (in the Mitra lab), Ryan Christensen and Yue Zhao for their suggestions and collegial spirit.

I will like to thank the Ewing Marion Kauffman Foundation for their fellowship in Life Sciences Entrepreneurship. This fellowship has introduced me to people and activities in the School of Law, Olin School of Business, and the Skandalaris Center for Entrepreneurship. It has afforded me the opportunity to be part of efforts to translate breakthroughs from the laboratory to the consumer.

A major portion of my doctoral career has been spent in collaborations with other scientists. Particularly my collaboration with Javier Irazoqui and his lab members, Orane Visvikis and Lyly Luhachack, have been extremely beneficial. They have been a pleasure to work with.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

C. elegans **as a model for studying transcription regulation**

C. elegans is a ~1mm transparent nematode which inhabits warm soil environments. An adult hermaphrodite has 959 somatic cells and an adult male has 1031 different cells. There are several attributes of C. elegans that make it an amenable system to studying transcription regulation: (1) It's non-variant cell lineage, which has been completely determined, has been very instrumental to developmental studies including: cell fates and differentiation stages (Sulston and Horvitz 1977; Kimble and Hirsh 1979; Sulston, Schierenberg et al. 1983; Kipreos 2005), the maternal to zygotic switch (Maduro, Broitman-Maduro et al. 2007), and silencing . (2) The multicellular organism is transparent and has several differentiated tissues (such as the pharynx, intestine, vulva, and pan-neurons) whose cells can be easily interrogated with fluorescence and other imaging modalities (Liu, Long et al. 2009). (3) These differentiated tissues (organogenesis) are similar to the development of organs in other complex organisms (Maduro 2006; Mango 2007; Mango 2009; Hobert 2010). Other conserved features of nematode biology that are used as models for other complex organisms include: components of the transcription machinery and mediator complexes (Casamassimi and Napoli 2007), aging (Antebi 2007; Baugh, Demodena et al. 2009), response to dietary restrictions (Baugh, Demodena et al. 2009), and reaction to pathogens (Irazoqui, Urbach et al. 2010). (4) The C. elegans genome has undergone little revision since first published. Current evidence suggests that the majority of transcriptional regulatory sequences are located in a relatively compact portion of the genome within about 2kb upstream of each

gene (Dupuy, Li et al. 2004; Zhao, Schriefer et al. 2007; Sleumer, Bilenky et al. 2009).

This reduced search space makes it possible to attempt to catalog all cis-regulatory

elements in this organism and study their effects on transcription. (5) Finally, C. elegans

has been extensively probed for the phenotypes of its ~20,000 genes. Studies have tested

genetic interactions between genes (Lehner, Crombie et al. 2006) and have used RNAi to

study the effects of knocking down about 85% of genes (Kamath and Ahringer 2003;

Lehner, Tischler et al. 2006).

Current challenges in discovering cis-regulatory elements

A long-standing problem in molecular genetics and genomics is the identification of all

the *trans*-acting factors (transcription factors, RNA-binding proteins, miRNAs, etc) that

regulate expression via their *cis*-sites in specific conditions and tissues. These factors and

sites are components of gene-regulatory networks (GRNs). Previous work has elucidated

aspects of these GRNs and explained biological mechanisms behind responses to specific

conditions (Arnone and Davidson 1997; Bolouri and Davidson 2002; Maduro and

Rothman 2002; Wenick and Hobert 2004; Hobert 2008; Gertz, Siggia et al. 2009). A

bottleneck in deriving these GRNs in C. elegans is the genome-wide identification of *cis*-

binding sites and the *in-vivo* specificity of *trans*-factors. There are an estimated 900 C.

elegans Transcription Factors (TFs) and more RNA-binding proteins. To directly probe

cis-bound regions (and indirectly infer TF specificities), some have employed chromatin

immunoprecipitation (ChIP) techniques to catalog transcription factor binding sites

(TFBS). ChIP is a laborious approach that is severely hampered by the following issues:

(1) the need to develop antibody tags for each factor; (2) the lack of adequate resolution

of binding sites due to sonication protocols; and (3) may not capture binding sites for

RNA-trans elements like RNA-binding proteins and miRNAs. As a result, there has been

slow progress in cataloging these sites using this method. The most comprehensive

collection of C. elegans ChIP-bound regions comes from the modENCODE project

which only has regions for 23 TFs (Gerstein, Lu et al. 2010).

Chromatin Immuno-Precipitation

Chromatin Imunno-Precipitation (ChIP) is a tool in molecular biology for probing the

DNA-bound regions for specific factors. This is accomplished by first crosslinking

factors in a cell to their bound regions using formaldehyde (or another agent). The cells

are then lysed to expose the genetic material and sonicated to break-up the bound

complexes to 300-1000bp fragments. The bound proteins are digested using proteases

and the extracted DNA is purified, amplified and   measured either by PCR methods,

hybridized cDNA chips (ChIP-CHIP) or next-generation sequencing (ChIP-Seq). These

steps are illustrated in the Figure 1.

Figure 1: Chromatin Immunoprecipitation procedure

An illustration of the major steps involved in chromatin immunoprecipitation. This figure was taken from Farnham (2009).
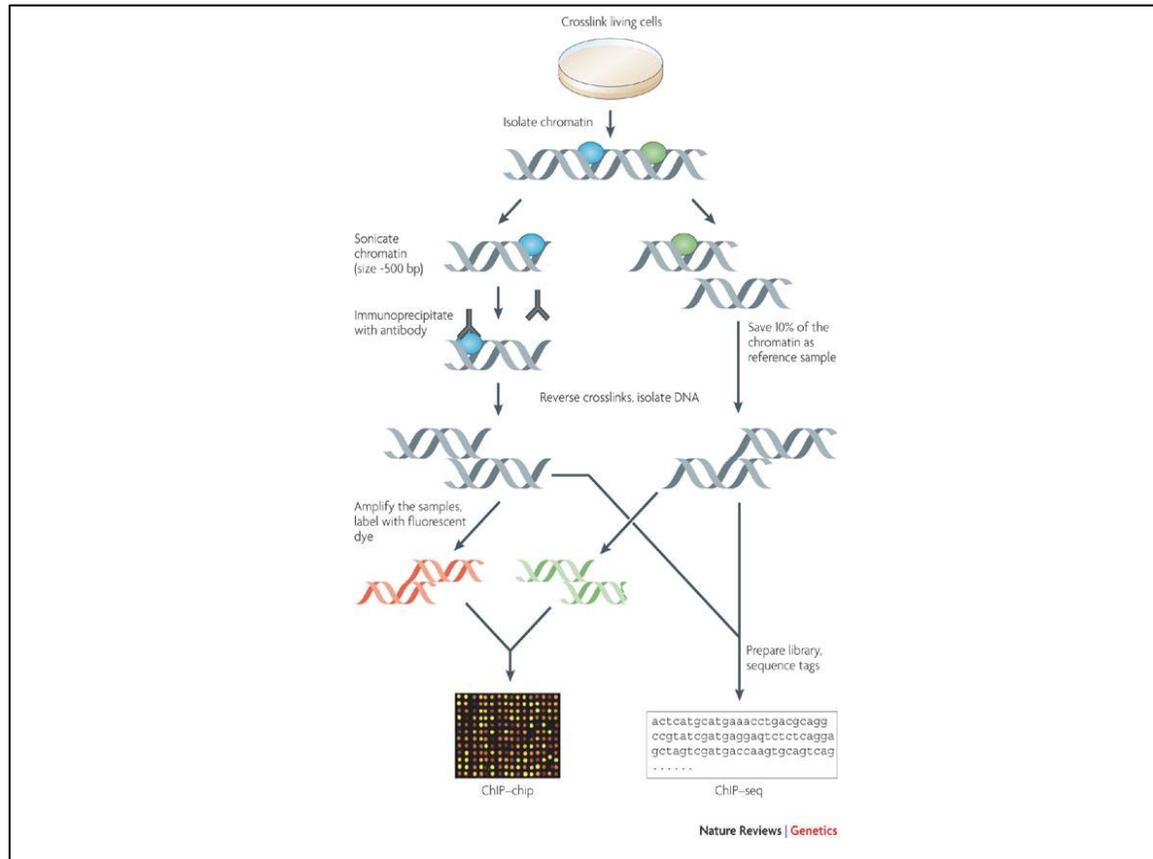
Motif-Finding

Another approach to discovering TFBS is to computationally search for frequently occurring similar sites within non-coding, regulatory regions of the genome (possibly conserved). These similar sites are then clustered into specificity models called motifs. This process is often referred to as motif-finding. These motifs represent putative binding specificities for corresponding TFs and other *trans*-factors.

Algorithms to recognize motifs in genomic DNA take one of two basic approaches. The *multiple gene, single species* approach recognizes motifs because they recur with few changes in the promoters of multiple genes within a single genome. These genes are usually the results of ChIP experiments. Therefore, they are known, or thought, to be co-regulated and expected to share a common motif. In contrast, the *single gene, multiple species* -- or *phylogenetic footprinting* – approach recognizes motifs in a single promoter region by their conservation across species, which is assumed to be greater than that of the surrounding background sequence (Gelfand 1999; McGuire, Hughes et al. 2000; McCue, Thompson et al. 2001; Panina, Mironov et al. 2001; Rajewsky, Socci et al. 2002; Frazer, Elnitski et al. 2003; Panina, Vitreschak et al. 2003; Marchal, De Keersmaecker et al. 2004). These methods work because binding sites are typically under selective pressure and therefore mutate more slowly than the surrounding sequence. Wang and Stormo (2003) combined these two approaches in their PhyloCon program, which runs not on individual DNA sequences but rather on alignments of orthologous promoter regions. In this paradigm, a motif is required both to recur across different promoters and to be conserved across species in each of its occurrences. Other tools that take a conceptually similar approach include (Qin, McCue et al. 2003; Jensen, Shen et al. 2005; Monsieurs, Thijs et al. 2006), all of which report results on bacterial promoters.

Resulting motifs can be represented as either: consensus sequence, count or frequency matrices, and logos. The following figure is an example these various representations for a set of similar sites (see figure 2).

Figure 2: Representations of cis-regulatory elements

SITES	CONSENSUS	POSITION SPECIFIC MATRIX	LOGO

```
SITES               CONSENSUS          POSITION SPECIFIC MATRIX          LOGO

ATGATAAGAT
TAGATAAGAA
ATGATAAGAT
AAGATAAGTT
AAGATAAGTT      WWGATAAGWW
AAGATAAGTA
TAGATAAGAA
TAGATAAGAA
ATGATAAGAT
AAGATAAGTT
ATGATAAGAT
AAGATAAGTT
AAGATAAGTA
TAGATAAGAA
```

A	7	7	0	14	0	14	14	0	7	7
C	0	0	0	0	0	0	0	0	0	0
G	0	0	14	0	0	0	0	14	0	0
T	7	7	0	0	14	0	0	0	7	7
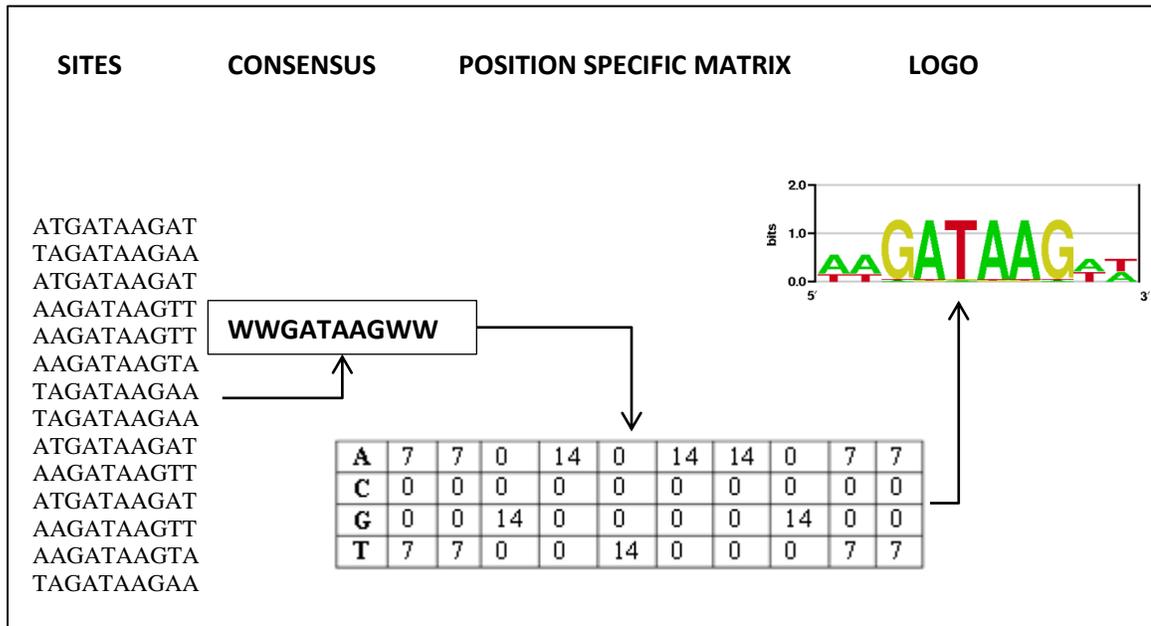
These motif models describe the binding specificity of a putative TF. Pioneering work by Berg and von Hippel (1987) introduced a statistical-mechanical framework for deriving this TF specificity/affinity. Their major assumptions, as is still adopted, were: (1) similar sites are bound by a TF with similar affinities and (2) contacting DNA-binding residues independently bind to nearby nucleotides. The first assumption, which is based on a natural selection argument, has been further observed in crystal structures of similar DNA sites bound to protein families with similar structures (Sandelin and Wasserman 2004). The second argument has also been shown to be generally acceptable. Most recently, Zhao and Stormo (2011) demonstrate that most protein specificity, ascertained by protein-bound microarrays (PBMs), can be adequately modeled with the independent nucleotide assumption, and find only a few exceptions where the pairwise interaction terms offer significant improvements.

Position Specific Scoring Matrices (PSSMs) are based on counts matrices. They consider both the observed and the expected background frequency of nucleotides in the genome by taking the logarithm of the ratios. The resulting 'binding energy' approximates the free energy required for a putative TF to bind to a DNA site (Stormo and Fields 1998). When summed over the genome, the resulting metric is proportional to the genome-wide occupancy of the TF (Granek and Clarke 2005; Bussemaker, Foat et al. 2007):

$$Occ(P_j, M_k) \propto \sum_{S_i \in P_j} e^{M_k \cdot S_i}$$

P_j is a specific promoter, M_k is a specific PSSM and S_i are all of the positions within the promoter. The score for any site with a given PSSM is $M_k \cdot S_i$ and is related to the logarithm of the probability of the site being bound by the TF whose specificity is represented by the PSSM.

Although they are less expensive than the ChIP techniques, computational results tend to be plagued by the following issues: (1) inability to identify the TF/factor that may bind via the discovered motifs; (2) Ineffective clustering of similar sites into motifs leading to reduced sensitivity, redundant motifs and unrealistic runtimes for organisms with larger regulomes (all non-coding sequences surrounding genes that harbor regulatory elements); (3) statistically significant sites may not be biologically functional because their similarity may just be artifacts of a relatively short evolutionary process. Additionally, they may be biologically insignificant because these computational approaches do not take into account the free concentration of TFs/factors in the cell. Since the concentration

of TFs is much less than the DNA sites in the cell, a 'perfect' TFBS may have the potential to be bound by a factor but may not be bound at a specific time or condition.

Most combined (multi-species, multi-gene) motif-finding approaches are generally characterized by three major steps: (1) Phylogenetic-footprinting to discover evolutionarily-conserved profiles; (2) Profile clustering to aggregate similarly conserved footprints observed at several locations in the genome, and (3) post-processing to remove redundant or unrealistic discovered elements. Due to the time and memory cost of the second and third stages most computational solutions for multicellular eukaryotic genomes have avoided whole-regulome searches (i.e. promoter, introns, and downstream regions of every gene) and only concentrate on specific regions near selected genes. In this thesis I present a new computational approach (Magma) that provides an adequate solution to these issues that plague motif-finding approaches and show tractable scaling times for higher-order organisms (Ihuegbu, Stormo et al. 2011).

Cis-Regulatory Modules

Gene expression is a non-linear function of multiple nearby cis-regulatory sites and the collection of trans-TFs bound to them. Arnone and Davidson (1997) showed that understanding gene expression requires not only the collection of individual TFBS, but clusters of these sites where multiple TFs may act to coordinately endow a specific regulatory mode (Moilanen, Fukushige et al. 1999; Gaudet and Mango 2002; Kirouac and Sternberg 2003; Natarajan, Jackson et al. 2004). These clustered sites, also called cis-regulatory modules, are comprised of TFBS that are nearby, perhaps tens of bases apart. The few known C. elegans modules collected in the Oreganno database of regulatory

8

elements average about 200 bases and are usually within 2kb upstream of the start site (Montgomery, Griffith et al. 2006; Griffith, Montgomery et al. 2008). In this thesis, I expand this collection of modules by predicting other modules in the promoter, intronic, and downstream regions near genes.

Determining differentially expressed genes

There are several different molecular biology techniques to ascertain the steady-state relative abundances of RNA molecules available at a snapshot within a single cell or tissue. The more popular techniques include: expression cDNA chips, RNA-Seq, and Promoter::GFP fusions. The earlier of these is cDNA expression microarrays in which microarray chips are pre-fabricated with spots of complimentary DNA matching regular intervals or particular regions of an organism's transcriptome (all genes, pseudo-genes, etc. that are transcribed in an organism's genome). As shown in Figure 3, these probes are hybridized with the extracted, labeled RNA sample from a cell in aqueous phase via hydrogen bonds between complimentary nucleic bases. After hybridization, unbound molecules are washed off the chip slide and the fluorescently labeled pairs are imaged with a scanner. The bound spots are illuminated and their relative abundance is estimated from the intensity of the fluorescence. More high density olignonucleotide libraries have since been fabricated on chips that provide even greater coverage of the transcriptome.

Figure 3: Expression Microarray

An illustration of a labelled RNA molecule extracted from cells of interest hybridizing to fixed complementary probes on a surface. This image and following legend were taken from Schulze and Downward (2001).

cDNA microarrays. Array preparation: inserts from cDNA collections or libraries (such as IMAGE libraries) are amplified using either vector-specific or gene-specific primers. PCR products are printed at specified sites on glass slides using high-precision arraying robots. Through the use of chemical linkers, selective covalent attachment of the coding strand to the glass surface can be achieved. Target preparation: RNA from two different tissues or cell populations is used to synthesize single-stranded cDNA in the presence of nucleotides labelled with two different fluorescent dyes (for example, Cy3 and Cy5). Both samples are mixed in a small volume of hybridization buffer and hybridized to the array surface, usually by stationary hybridization under a cover-slip, resulting in competitive binding of differentially labelled cDNAs to the corresponding array elements. High-resolution confocal fluorescence scanning of the array with two different wavelengths corresponding to the dyes used provides relative signal intensities and ratios of mRNA abundance for the genes represented on the array. **b**, High-density oligonucleotide microarrays. Array preparation: sequences of 16–20 short oligonucleotides (typically 25mers) are chosen from the mRNA reference sequence of each gene, often representing the most unique part of the transcript in the 5'-untranslated region. Light-directed, *in situ* oligonucleotide synthesis is used to generate high-density probe arrays containing over 300,000 individual elements. Target preparation: polyA$^+$ RNA from different tissues or cell populations is used to generate double-stranded cDNA carrying a transcriptional start site for T7 DNA polymerase. During *in vitro* transcription, biotin-labelled nucleotides are incorporated into the synthesized cRNA molecules. Each target sample is hybridized to a separate probe array and target binding is detected by staining with a fluorescent dye coupled to streptavidin. Signal intensities of probe array element sets on different arrays are used to calculate relative mRNA abundance for the genes represented on the array.

cDNA microarray   High-density oligonucleotide microarrays
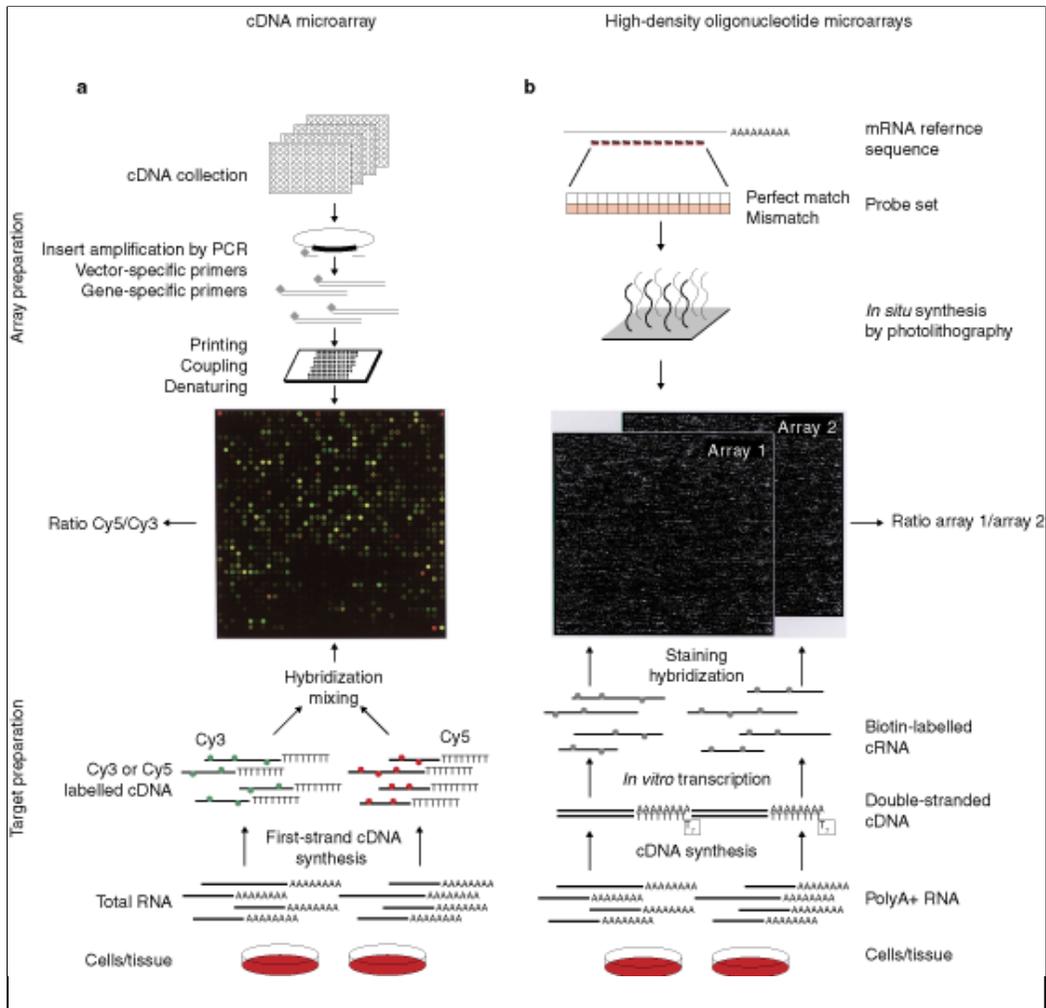
A recent, more throughput method is RNA-Sequencing coupled with Next-Generation

Sequencing. In this approach, RNA molecules are extracted and prepared with adapters

and primers for sequencing. Next-Generation Sequencing platforms, such as Solexa, are

capable of quickly generating millions of reads tracing RNA molecules using a library of

cDNA molecules. After extension and sequencing, reads are aligned to a reference

genome and clustered.. The normalized number of reads aligned to each gene model is

proportional to the abundance of the transcript. This process is demonstrated in Figure 4

which is adapted from Figure 1a by Mortazavi, Williams et al. (2008).

Figure 4: RNA-Sequencing

A description of the major steps involved in RNA-Sequencing. This image was adapted from Mortazavi, Williams et al. (2008).
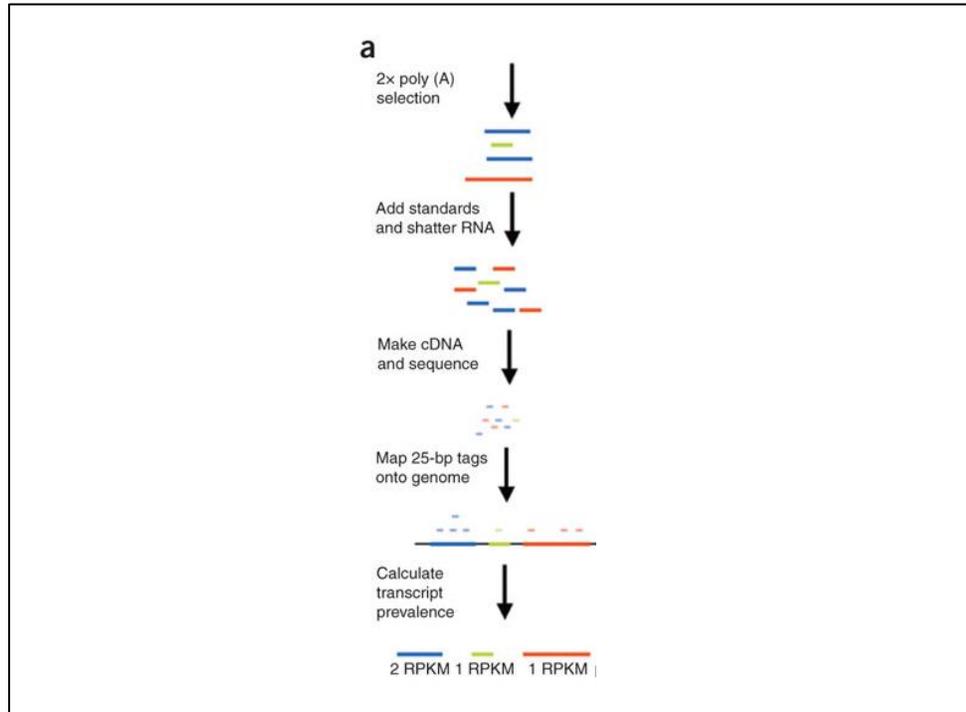
Reporter promoter∷GFP fusions offer the lowest throughput of these methods, but they also offer the greatest gene-specificity as they allow for observation of gene activation in its *in vivo* temporal and spatial contexts. In this method, promoters of interest (or the entire gene body as seen in Figure 5) are fused to a green fluorescent protein gene (GFP) and transformed or transfected into the organism of interest. These reporter genes are used to monitor the activation of specific genes at a time during development or in response to stimuli. Additionally they can then be used to filter genes during development as done in FACs sorting and other screening assays (Boulin, Etchberger et al. 2006; Koo, Kim et al. 2007).

Figure 5: Reporter:GFP constructs

A brief description of the general steps involved in designing a C-terminal reporter:GFP fusion, taken from (Boulin, Etchberger et al. 2006)Primers A and B amplify the genomic region (amplicon #1). Primer B adds a 24 bp overlap in frame to the GFP coding region. Primers C and D amplify the reporter gene (e.g., GFP) and 3' UTR (amplicon #2). Primers A* and D* are used to fuse amplicon #1 and amplicon #2 (gray box indicates 24 bp sequence overlap). The resulting fusion product (amplicon #3) can be directly injected into *C. elegans* without purification.
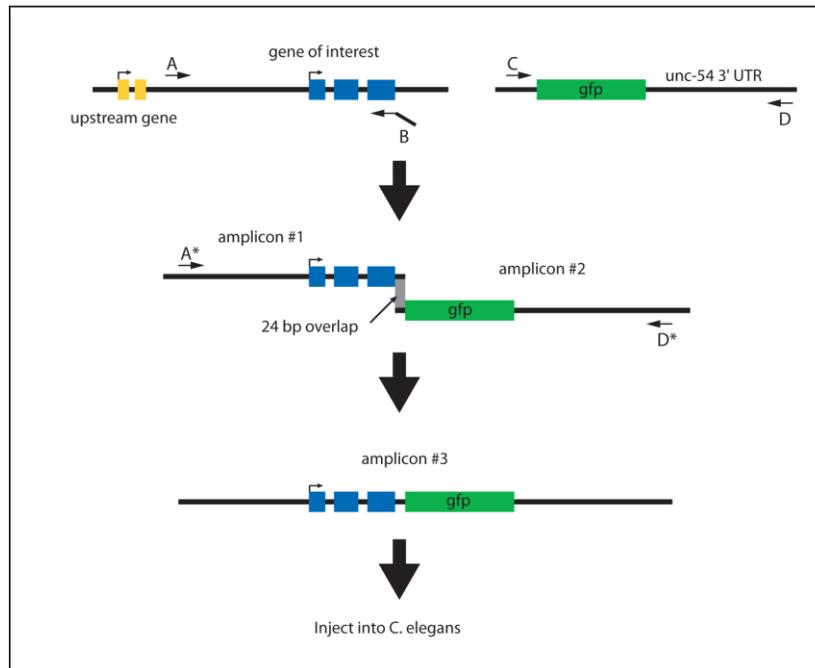
Associating putative cis-regulatory elements to expression

Ongoing challenges in finding cis-regulatory elements on a genome-wide scale for higher- eukaryotes has also made it difficult to associate *cis*-elements to differential expression on a genome-wide scale. Earliest methods to associate cis-regulatory elements to expression, involved searching for the presence of similar sites near gene modules -- clusters of similarly expressed genes (Eisen, Spellman et al. 1998). This spurred the use of Gibbs sampling techniques to find over-represented or enriched motifs in the

promoters, inferring causation between the presence of these sites and the expression level (Spellman, Sherlock et al. 1998; Tavazoie, Hughes et al. 1999). Beginning with Segal, Dahlquist et al. (2003) and Beer and Tavazoie (2004), mRNA expression levels were predicted as simply the average from all genes in a set after clustering based on presence of similar combinations of motifs.

More recent quantitative approaches begin by using a PSSM to define a metric that is proportional to the occupancy of the TF. Bussemaker, Li et al. (2001) introduced sequence-based linear regression techniques to model the gene expression levels based on the occupancy of independent multivariate regulatory motifs. In their method they used a forward variable selection to include orthogonal motifs discovered in the promoter sequences that explained the most variation in the observed expression levels. Since then others have adopted similar regression techniques (Foat, Houshmandi et al. 2005; Foat, Morozov et al. 2006). Keles, van der Laan et al. (2002) extended this approach by introducing terms to capture the location bias of regulatory motifs in the promoter and selecting significant variables using feature selection. Others have also used expectation maximization methods to infer TF promoter-specific affinities and regulatory effects (Nachman, Regev et al. 2004; Tanay and Shamir 2004). These quantitative techniques offer more statistical rigor than other qualitative (enrichment) methods but they require more parameters and make more predictions than are actually tested (such as the relative weight of motifs in predicting gene expression).

Overview of this thesis

In this work, I describe initial efforts to comprehensively discover cis-regulatory elements in promoters of *C. elegans* genes using PhyloNet (Zhao, Ihuegbu et al. 2011). I then furthered this to enable efficient scaling for the rest of the regulome and for higher-order organisms (with larger non-coding regions). The resulting software tool, Magma (Multiple Aligner of Genomic Multiple Alignments), is an efficient method for discovering conserved elements that recur several times in a eukaryotic genome (Ihuegbu, Stormo et al. 2011). Although it is motivated by PhyloNet, it differs in important ways that make it much faster and somewhat more sensitive. Consequently, for the first time, intergenic, UTR, and intronic elements are discovered in the C. elegans regulome. Magma efficiently clusters millions of sites into motifs and is fairly sensitive in recovering known regulatory elements and modules as well as their associations to expression. Furthermore, in collaboration with Javier Irazoqui and Orane Visvikis, I describe a novel discovery in which sites that comprise a Magma-discovered motif, which represents a binding preference for HLH-30, are bound by HLH-30 in response to S. aureus infection. When this TF is knocked out, C. elegans animals have a significantly increased susceptibility to the pathogen.

This simple approach to finding and relating cis-regulatory elements to expression (via enrichments) excludes several important determinants such as: (1) Chromatin/histone structure, (2) Copy number variation, and (3) binding-site turnover. Nevertheless, many of our findings correspond to known *trans*-factors that have been previously implicated with specific conditions/tissues. Furthermore, I predict several novel mechanisms of

regulation implicating known factors with new conditions and/or new regions of

regulation.

Chapter 2: Conserved Motifs and Prediction of Regulatory Modules in *Caenorhabditis elegans*[1]

Abstract

Transcriptional regulation, a primary mechanism for controlling the development of multicellular organisms, is carried out by transcription factors (TFs) that recognize and bind to their cognate binding sites. Our understanding of transcriptional regulation in *C. elegans,* which TFs bind to which sites and regulate which genes, is still very limited. To expand our knowledge about the *C. elegans* regulatory network, we performed a comprehensive analysis of the *C. elegans*, *C. briggsae* and *C. remanei* genomes to identify regulatory elements that are conserved in all genomes. Our analysis identified 4959 elements that are significantly conserved across the genomes and that each occur multiple times within each genome, both hallmarks of functional sites. Our motifs show significant matches to core promoter elements, known TF binding sites, splice sites and poly-A signals as well as many putative regulatory sites. Many of the motifs are significantly correlated with various types of experimental data including gene expression patterns, tissue specific expression patterns and binding site location analysis as well as enrichment in specific functional classes of genes. Many can also be significantly associated with specific TFs. Combinations of motif occurrences allow us to predict the location of cis-regulatory modules and we show that many of them significantly overlap experimentally determined enhancers. We provide access to the predicted binding sites, their associated motifs and the predicted cis-regulatory modules across the whole genome through a web-accessible database and as tracks for genome browsers.

Introduction

The development of an organism is largely controlled by transcriptional regulation which determines where and when every gene is expressed. A first step toward the understanding of how genomic DNA controls the development of an organism is to understand the mechanisms that control differential gene expression. Transcriptional regulation is carried out by transcription factors (TFs) via their binding to specific DNA sequences. Binding sites of TFs can be represented as consensus sequences but position weight matrices (PWM) provide a more quantitative description of the specificity of a TF (Stormo 2000). Currently our knowledge of the TFs and their binding sites is very limited. For example, the human genome has greater than 2000 predicted TFs (Lander, Linton et al. 2001) but only a few hundred have quantitative models of their specificity, primarily based on computational tools that have been developed to facilitate the identification of PWMs for TFs (reviewed in GuhaThakurta 2006). Furthermore, although computational methods can successful identify binding sites that are bound by a particular TF *in vitro*, most of the predicted binding sites are not functional *in vivo* (Whittle, Lazakovitch et al. 2009; Li, Thomas et al. 2011). Previous studies have shown that TF binding sites tend to cluster together to direct tissue/temporal-specific gene expression (Kirchhamer, Yuh et al. 1996; Arnone and Davidson 1997). These clusters of binding sites that regulate expression are referred to as cis-regulatory modules (CRMs). Clustering of TF binding sites, along with phylogenetic conservation and other measures of "regulatory potential", have been widely used in computational prediction of CRMs and is a more reliable indicator of *in vivo* regulatory function of DNA sequences (Kolbe, Taylor et al. 2004; Wasserman and Sandelin 2004; King, Taylor et al. 2005; Blanchette,

19

Bataille et al. 2006; Sinha, Liang et al. 2006; Taylor, Tyekucheva et al. 2006; Ferretti, Poitras et al. 2007) .

C. elegans has been an important model organism for studying development and was the first metazoan with a completely sequenced genome (Consortium. 1998). While a few promoters have been studied in detail (Krause, Harrison et al. 1994; Gaudet and Mango 2002; Ao, Gaudet et al. 2004; McGhee, Sleumer et al. 2007; McGhee, Fukushige et al. 2009), most transcriptional regulatory interactions remain unknown. Recently projects have been undertaken to gain a more comprehensive view of which TFs regulate which promoters using experimental approaches to identify their interactions directly (Deplancke, Mukhopadhyay et al. 2006; Celniker, Dillon et al. 2009; Gerstein, Lu et al. 2011) but those are still in early phases. A complementary approach is to identify non-coding segments of the genome that are conserved across species and are likely to contain regulatory elements (reviewed in Wasserman and Sandelin 2004). There are several previous works on regulatory motif prediction in *C. elegans* (GuhaThakurta, Palomar et al. 2002; Ao, Gaudet et al. 2004; Gaudet, Muttumu et al. 2004; GuhaThakurta, Schriefer et al. 2004). However, those focus on sets of genes that are expressed under specific conditions or in specific tissues. A recent report compared eight nematode species and identified regions conserved in *C. elegans* and at least three other species for over 3800 genes which are catalogued in their cisRED database (Sleumer, Bilenky et al. 2009). In this paper we performed a genome-wide cis-regulatory element identification using PHYLONET (Wang and Stormo 2005), which systematically identifies phylogenetically conserved motifs that also occur multiple times throughout the genome and are likely to define a network of regulatory sites for a given organism. The first step of this approach

20

is similar to that used for cisRED, identifying segments conserved across multiple species, but then it further compares all such conserved regions to each other to identify those associated with multiple genes. Applying PHYLONET on 2kb intergenic regions from the genomes of *C. elegans*, *C. briggsae* and *C. remanei* leads to the identification of cis-regulatory elements from various functional categories. We identified core promoter elements, TF binding sites, splicing sites, poly-A signals as well as binding sites of non-TF proteins. In addition, for each regulatory element, PHYLONET identified a set of genes which are potentially regulated by the motif. Gene functional enrichment and expression coherence analysis under several conditions provide strong support that most of the motifs are functional elements that are responsible for the regulation of the target genes. The instances of these predicted cis-regulatory elements along the promoter sequences are highly clustered. Based on this observation we developed a program, CERMOD, to predict new CRMs. Comparison between the predicted modules with experimentally characterized modules shows high sensitivity with 83.2% (124/149) of experimentally characterized modules. For genes with experimentally determined CRMs 47.9% (135/282) of our predicted modules are located within experimentally defined regions. This is a lower bound of predictive accuracy because many of our predicted modules could be real but are located within promoter regions that haven't been tested.

Material and Methods

Genome Sequences

The chromosomal sequence and the gene structures of *C. elegans* (Consortium. 1998) (WS170) and *C. briggsae* (Stein et al. 2003) genome are downloaded from the Wormbase ftp-site (ftp://ftp.wormbase.org/pub/wormbase/genomes/). These were then used to obtain -2000 to -1 upstream region sequences. *C. remanei* sequence and annotation were produced by the Genome Sequencing Center at Washington University School of Medicine in St. Louis and were obtained from http://genome.wustl.edu/pub/organism/Invertebrates/Caenorhabditis_remanei/.

Identification of orthologs of *C. elegans* genes

C. briggsae orthologs of *C. elegans* genes were obtained from WormBase ftp://ftp.wormbase.org/pub/wormbase/datasets/stein_2003/orthologs_and_orphans/orthologs.txt.gz. To identify *C. elegans* orthologous genes in the *C. remanei* genome, we used the NCBI BLAST program (version 2.0) (Altschul et al. 1990) to compare all annotated protein coding gene sequences in the *C. remanei* genome with that in the *C. elegans* genome. Two genes are defined to be orthologous if all of the following three conditions are met: (i) their protein sequences are reciprocal best BLASTP hits between two genomes, (ii) the BLASTP E-value is lower than 1E-10, and (iii) the BLAST alignment covers ≥ 60% of the length of at least one sequence. The promoter sequences (defined as -1 to up to -2000 intergenic sequences upstream of translational start site ATG) of all

genes in the orthologous gene set that contain both *C. briggsae* and *C. remanei* orthologs

of *C. elegans* gene were retrieved. Each sequence group of orthologous genes forms a

data entry. For *C. elegans* genes that are in operons (Blumenthal et al. 2002), we only

considered the first genes in the operons.

Motif identification and consolidation

We used PHYLONET, a program that systematically identifies phylogenetically

conserved motifs and defines a network of regulatory sites for a given organism (Wang

and Stormo 2005), to search for conserved regulatory elements. PHYLONET was run

with options s= 1, iq = 20, id = 20 and pf = 10. Up to 10 predicted cis-regulatory

elements are reported for each intergenic region. Cis-regulatory elements are represented

by position weight matrices (PWMs)(Stormo 2000) and each matrix is associated with a

set of genes that are potentially regulated by this element (gene cluster).

The initial motifs generated by PHYLONET are redundant because each gene is

used as a query and different gene queries can generate very similar motif profiles and

target gene clusters. To remove redundancy of the whole genome motif profile set, we

used the average log likelihood ratio (ALLR) statistic (Wang and Stormo 2003) to

determine the similarity between motif profiles. ALLR statistics are implemented in

MatAlign-v4a (Wang and Stormo, unpublished). Similarity of two motif profiles is

determined by the ALLR scores of each pair of motif profiles and the length of the

aligned part of the two motifs. To determine the best parameters for clustering PWMs,

we analyzed matrices in the TRANSFAC database (Matys, Fricke et al. 2003).

TRANSFAC version 10.2 contains 811 PWMs, 540 of which have known binding factors

that are classified at the family level. PWM similarity is measured with two parameters: ALLR score and OLAP score, which is the percentage of the two PWMs that overlap. At each ALLR and OLAP score cutoff value, we compare each of the 540 matrices with all of the others to determine the score distributions. From this information we calculate sensitivity and specificity for classifying each PWM into the correct family at each ALLR and OLAP cutoff value. Our results suggest that ALLR > 6.57, OLAP > 68.1% gives the best specificity. For all PHYLONET output matrices, the best one is picked first (the one with the highest Total ALLR score in PHYLONET output). Then it is compared with the rest of the matrices using ALLR statistics. Any matrix that appears redundant to the chosen matrix is removed. Then the second best one is picked and the process is repeated until all the matrices have been analyzed.

Calculation of functional enrichment of target genes sharing the same motif profile.

We tested the functional enrichment of target genes of each motif profile based on Gene Ontology (GO). GO terms and annotations of *C. elegans*, were downloaded from WormBase. All genes sharing the same GO term are clustered. Based on GO term hierarchies, we added all genes in the children GO terms to the current GO term gene cluster. The cumulative hyper-geometric distribution (Tan et al. 2005; Tavazoie et al. 1999) is used to calculate the P-value of observing the number of genes associated to a motif profile and enriched in a particular GO term.

Calculation of microarray expression profile coherence

Microarray expression profiles are downloaded from the Gene Expression Omnibus (GEO). Expression Coherence (EC) score and threshold distance (D) were calculated as described (Pilpel et al. 2001). We define the gene clusters to have significant expression coherence when their P-value < 0.05 after correction for multiple tests.

Cis-regulatory module identification

To identify DNA regions enriched for predicted motifs, we first identify all predicted sites for all the motifs using Patser (Hertz and Stormo 1999) using default cutoff scores. Then we calculate the average number of binding sites per position in the sequence and Z score for each position. We identify those peak positions that have a Z score ≥ 3.09 (corresponding to p-value = 1.0E-3). For each peak position, we extend it in both 5' and 3' direction if the next Z score > 0 position is less than 30 bp away (the longest motif length). Peak positions used in a previous extension step are not extended.

Results and Discussion

This section is divided into four subsections:

I: Overview of the conserved motifs identified by PhyloNet

II: Correspondence between the motifs and several different types of experimental data to
    assess their likely functions

III: Using the motifs to predict cis-regulatory modules across the entire intergenic regions

of C elegans, and an assessment of the accuracy of those predictions

IV: A description of the database of exemplar sites and motifs and of the genome browser

that facilitate access to the sites, motifs and module predictions.

I. Overview of the conserved motifs identified by PhyloNet

To systematically identify conserved elements in *C. elegans*, we used the genome

sequences from *C. briggsae* and *C. remanei*. We obtained 11,860 *C. briggsae* orthologs

and 12,466 *C. remanei* orthologs for 16,544 *C. elegans* genes. Some *C. elegans* genes are

organized as operons and genes in operons share a common promoter sequence that

allows coordinated expression of the genes. After removing the distal genes in operons,

as annotated in Wormbase (http://www.wormbase.org/), 10,491 and 11,064 of the *C.*

elegans genes have *C. briggsae* and *C. remanei* orthologs, respectively. 9,356 genes that

have both *C. briggsae* and *C. remanei* orthologs were used for further analysis.

Current evidence indicates that *C. elegans* regulatory regions are fairly compact and

most known regulatory elements occur within 2kb upstream of the coding region of the

gene (Dupuy, Li et al. 2004; Zhao, Schriefer et al. 2007; Sleumer, Bilenky et al. 2009).

We retrieved up to 2kb upstream promoter sequences for all of the genes with orthologs

in each species. Each *C. elegans* gene and its orthologs form a data entry which contains

three promoter regions. For each data entry, PHYLONET (Wang and Stormo 2005) was

applied to query the database and up to 10 most significant predicted motifs, represented

as position weight matrices (PWMs) (Stormo 2000), were obtained for further analysis.

Because of the greedy and reciprocal nature of the PHYLONET algorithm, where each promoter serves as the query for a BLAST-like alignment to every other promoter, these initial predicted motifs in the PHYLONET output files are highly redundant. We took two steps to consolidate predicted motifs. The first step compares matrices in each query output file to consolidate matrices that significantly overlap. This step results in a total of 36953 PWMs, an average of 3.95 PWMs for each C. elegans promoter. This set of sites is called the exemplar sites, those identified by PHYLONET as being conserved in the three species and significantly similar across multiple genes. From the initial set of nearly 20Mbp in candidate regions from C. elegans, the exemplar motifs cover a total of 3,695,282bp, which is about 18% of the intergenic regions considered.

The second step is to consolidate PWMs based on motif similarity to generate the final set. This step is challenging because our goal is to find cis-acting regulatory motifs that correspond to all of the trans-acting regulatory factors, but there is not a simple one-to-one relationship between them. One complication is that TFs from the same structural family often bind to highly similar DNA target sequences (Luscombe, Austin et al. 2000) and it can be difficult to separate sites for different TFs based on the conserved motifs alone. Several computational approaches have been developed to quantify similarities between PWMs (Wang and Stormo 2003; Kielbasa, Gonze et al. 2005; Schones, Sumazin et al. 2005) and to use this information to classify the structural class of mediating TFs for novel motifs (Sandelin and Wasserman 2004; Kielbasa, Gonze et al. 2005; Schones, Sumazin et al. 2005; Narlikar and Hartemink 2006). Mahony et al. (Mahony, Auron et al. 2007) evaluated various comparison metrics and alignment algorithms for comparing PWMs. We use the average log-likelihood ratio (ALLR) (Wang and Stormo 2003) to

cluster motifs into distinct sets. Although Mahony et al. (Mahony, Auron et al. 2007) did

not find ALLR to be the best statistic for assigning motifs to TF structural classes, our

most challenging goal is to distinguish similar motifs from the same class, for which

ALLR is well suited.

Our stringent criteria (see MATERIALS AND METHODS) allow only very similar

motifs being clustered together. This gives us confidence that we have not merged motifs

for different TFs, but has the disadvantage that we may have several distinct PWMs

remaining for the same TF. This is certainly the case as the second consolidation step

leaves us with 4959 distinct motifs with lengths between 5 and 30 bases, many more than

the proposed number of about 940 *C. elegans* TF genes (Reece-Hoyes et al. 2005). These

motifs cover 3,442,144 bp and have an average length of about 15bp. We find other types

of known motifs besides TF binding sites (see below), but in addition the motifs probably

contain sites for combinations of TFs which we have not separated into distinct sub-sites.

These consolidated PWMs are all very significant ($p<10^{-10}$), and each is associated with a

set of genes that are potentially regulated by this motif. Each consolidated PWM is

associated with a set of exemplar sites and a gene list. The gene lists range from 3 to

7724 genes. We expect the exemplar sites for each PWM to be an incomplete set of

binding sites for the associated factor because less than half of the C. elegans genes are

used in our initial promoter set and because, even for orthologous genes, some sites will

not be conserved across the different species. We can use the PWMs to predict other

potential binding sites for the associated factor. These predicted sites should provide a

more comprehensive list of binding sites, and regulated genes, for each PWM, but will

likely also include some false predictions.

II. Correspondence between the motifs and several different types of experimental data

Intergenic regions contain different kinds of regulatory elements. We are particularly interested in TF binding sites involved in controlling gene expression, but other elements are also obtained in our set of conserved motifs. Figure 6 shows three different classes of conserved elements that emerge from this analysis. The PWM H01M10.2.1 is likely to represent the binding motif for the transcription factor NFI-1 based on several types of evidence: 1) it is highly similar to the documented NFI-1 binding site (Whittle, Lazakovitch et al. 2009) and the vertebrate NF-1 binding site (TRANSFAC AC number: M00056) 2) the gene cluster associated with the H01M10.2.1 matrix is significantly enriched for the known NFI-1 target genes (Whittle, Lazakovitch et al. 2009) ($p<10^{-14}$). 3) the gene cluster associated with the H01M10.2.1 matrix is significantly enriched for genes that are expressed in pharynx ($p<3x10^{-5}$) and body wall muscle ($p<7x10^{-3}$) which is consistent with observed NFI-1 expression in C. elegans (Lazakovitch, Kalb et al. 2005; Lazakovitch, Kalb et al. 2008). 4) H01M10.2.1 is significantly correlated to NFI-1 ChIP samples ($p<10^{-6}$; t-value ~ 15.6). 5) the gene cluster associated with the H01M10.2.1 matrix is significantly enriched for GO terms that are consistent with NFI-1's function. A second type of element we obtain is a core promoter motif such as the TATA-box (Figure 6). K09B3.1.8 matrix is very similar to the TRANSFAC TATA box PWM (M00216) and, unlike most transcript factors binding sites, it is significantly biased in its location and its orientation. It is significantly over-represented near the translational start site ATG, in positions between 21-40 ($p<10^{-10}$) and $41 - 60$ ($p<10^{-10}$) nucleotides upstream of the ATG. It is also preferentially located on the + strand ($p < 10^{-3}$) as expected for a core

promoter element. A third type of element we find are those related to RNA processing. In *C. elegans* more than half of pre-mRNAs are subject to SL1 trans-splicing (Blumenthal and Steward 1997). The trans-splice site consensus on the pre-mRNAs is the same as the intron 3' splice site consensus. Y94H6A.1.1 matrix is significantly similar to the *C. elegans* trans-splice/3' splice signal. It is significantly over-represented near the translational start site ATG (Figure 6). However, different from the TATA box, it is preferentially located between 0 to 20 nucleotides upstream of ATG ($p<10^{-20}$). Trans-splicing occurs close to the start codon in *C. elegans* with 49% of transcripts analyzed containing a spliced leader sequence within 10 nucleotides of the initiator AUG (Lall, Friedman et al. 2004). In addition, Y94H6A.1.1.matrix is preferentially located on the + strand ($p < 10^{-3}$) as expected for a splicing signal. We did not find a motif that represents the 5' splice signal which is consistent with the presence of 3' but not 5' splice signal in front of ATG in the case of trans-splicing.

Besides identifying PHYLONET PWMs that correspond to motifs for known factors as described above, we can assess whether the genes associated with any PWM are significantly correlated with specific biological assays. In the following sections we consider data from four different approaches: 1) transcription factor binding data, such as ChIP-chip and ChIP-seq experiments that identify binding locations for specific TFs; 2) expression data, such as microarrays that measure gene expression patterns under specific conditions or specific genetic backgrounds; 3) tissue specific expression patterns of genes using GFP-fusions; 4) enrichment for specific classes of genes using gene ontology (GO) classifications for genes. Significant correlations between the genes selected by any of

those methods and genes associated with one of our PWMs provides supporting evidence that the PWM represents a regulatory motif.

II.1 Location analysis

Regions in the genome where TFs bind *in vivo* can be determined experimentally by expressing tagged *C. elegans* TFs that then are cross-linked to chromatin and their locations determined by either array hybridizations (ChIP-chip) or sequencing (ChIP-seq). We compare those experimentally determined binding locations to the predicted occupancy for each PWM on each promoter. The predicted occupancy is calculated by scoring each position in the promoter with the PWM and summing the exponentiated scores:

$$Occ(P_j, M_k) \propto \sum_{S_i \in P_j} e^{M_k \cdot S_i}$$

where P_j is a specific promoter, M_k is a specific PWM and S_i are all of the positions within the promoter. The score for any site with a given PWM is $M_k \cdot S_i$ and is related to the logarithm of the probability of the site being bound by the TF whose specificity is represented by the PWM (Stormo 2000; GuhaThakurta, Schriefer et al. 2004; Granek and Clarke 2005; Chang, Nagarajan et al. 2006). The proportionality constant that relates this occupancy score to the true occupancy of the promoter is unknown but is not needed because we use a correlation coefficient to compare the occupancy score to the experimental determinations of binding locations.

A total of 57 binding assays, including array hybridizations (ChIP-chip, 39 samples) and sequencing (ChIP-seq, 18 samples), were obtained from the GEO database (Barrett, Troup et al. 2009). Individual samples are further processed where appropriate, such as comparing a specific ChIP-chip array to its control array or to different time points in a time series experiment. Thus 51 experiments were used in the analysis and a total of 794 motifs have a predicted occupancy that is significantly correlated ($p<0.01$, or t-value > 6.02 after correcting for multiple tests) with at least one of the 51 different processed samples (see Supplemental material: "Motifs Significant in Location Analysis" for the complete list). An example is H01M10.2.1 which is significantly correlated to NFI-1 ChIP samples ($p<1\times10^{-6}$; t-value ~ 15.6) (Whittle, Lazakovitch et al. 2009). Using ChIP-Seq Whittle et al identified 55 genes that passed a strict cutoff for binding. The motif they identified in 49 of the 55 bound regions is nearly identical to our motif H01M10.2.1 and also to a previously reported motif for vertebrate NFIs (Figure 6). Of those 55 genes, 36 were included in the promoter sets analyzed by PHYLONET and 32 of them had NFI-1 binding sites identified by Whittle et al within the 2kb upstream regions of our study. The PHYLONET PWM H01M10.2.1 contains 22 exemplar sites from that set of 32 reported NFI-1 sites ($p<10^{-14}$).

II.2 Expression analysis

The same predicted occupancy scores for each PWM and each promoter can be compared to expression data to determine if motif occurrences are significantly correlated with expression patterns. A total 1197 expression samples were obtained from the GEO database and further processed where appropriate, as in the location analysis. A total of

797 motifs are significantly correlated (p<0.01 after correction for multiple tests) with at least one of 850 different processed samples (see Supplemental material: "Motifs Significant in Microarray Expression Analysis" for the complete list).

In the location analyses described above, we uncovered associations between specific motifs and the specific proteins that were immunoprecipitated. This lets us infer that the motif represents the binding specificity of the protein, or perhaps another protein that is tightly coupled to the one that is precipitated. In the expression analysis we identify motifs that are associated with genes whose expression changes under different conditions, genetic backgrounds or at different times or different tissues during development. We can hypothesize that the motifs represents the binding sites for some proteins responsible for these changes in expression, but the identity of the proteins is usually unknown. However, in some cases we find the same motif identified in the location analysis and the expression analysis which suggests that the specific protein acts through the identified motif to control the expression of the regulated genes. We find 424 such motifs that are significant in both datasets.

Although our collection of expression microarrays do not include any records in which NFI-1 mutants were probed for genome-wide expression, previous work that suggest NFI-1 is critical for wild type adult lifespan (Lazakovitch, Kalb et al. 2005). We observe significant correlations between occupancy scores for H01M10.2.1 on nearby genes and their expression changes in age-related micorarray records (p<0.01). Additionally, another discovered motif C18D1.B.5 is similar to the core portion of previously discovered Motif Enriched on X (MEX) motif which Jans et al show to be a component of the Dosage Compensation Complex (DCC) (Jans, Gladden et al. 2009).

Interestingly, our location analysis results show that C18D1.B.5 is correlated with the Chip-Chip results for the DCC subunits (DPY27: $p<10^{-16}$; SDC-2: $p<10^{-16}$; SDC-3: $p<10^{-14}$; MIX-1: $p<10^{-3}$; HTZ-1: $p<10^{-16}$). Additionally, our expression analyses show that C18D1.B.5 matrix is also correlated with XO vs. XX-WT expression studies ($p<10^{-16}$).

Co-regulated genes often have similar expression profiles under different conditions. We can thus evaluate the likelihood of a motif being biologically meaningful by the coherence of the expression profiles of all the target genes associated with the motif. We used the expression coherence score (Pilpel, Sudarsanam et al. 2001) to measure the overall similarity of the expression profiles of all the target genes of a given predicted motif in several different conditions. The NCBI GEO database contains 9 datasets that studied C. elegans gene expression under different conditions or at different time points and therefore are suitable for expression coherence analysis. The 9 data sets are PAL-1 network (GDS1319), Hypoxia response (GDS1379), TOM1/UNC-43 (GDS1786), Twist over-expression (GDS2463), lin-35 null mutant at various stages of development (GDS2751), Aging time course (GDS583), heat stress time course (GDS584) Germline development (GDS6) and daf-2 mutant expression profiling (GDS770). Using a stringency cutoff of $p<0.05$ after correction for multiple tests we determined that 682 (13.75%) exemplar gene sets exhibit similar expression patterns in at least one experimental condition, suggesting a regulatory function of that associated PWM (see Supplemental material: "Motifs Significant in Expression Coherence Analysis" for the complete list). H01M10.2.1 matrix associated genes, described above as associated with NFI-1 binding, have significant expression coherence in both hypoxia response experiment and heat stress time course experiment. Interestingly NF1 in Drosophila has

similar functions of regulating life span as that of C. elegans NFI-1 and flies over-expressing *NF1* had increased life spans, improved reproductive fitness, increased resistance to oxidative and heat stress in association with increased mitochondrial respiration and a 60% reduction in ROS production (Tong, Schriner et al. 2007)

C47A10.6.1 is similar to the heat shock element (HSE) identified in the promoters of the genes that were consistently up-regulated 1 and 4 hr after heat shock (GuhaThakurta, Palomar et al. 2002). Genes associated with C47A10.6.1 have significant expression coherence in both hypoxia response experiment and heat stress time course experiment but not in any other experiment. F01G4.4.5 is similar to the heat shock associated site (HSAS) identified in the same study as HSE (GuhaThakurta, Palomar et al. 2002). Similarly, genes associated with F01G4.4.5 have significant expression coherence in both hypoxia response experiment and heat stress time course experiment but not in any other experiment.

II.3 Tissue-specific expression patterns

1,882 *C. elegans* transcripts (~10% of the genome) have classified expression patterns in 88 different spatial-temporal patterns between the larval and adult stages (Hunt-Newbury, Viveiros et al. 2007). Ignoring the developmental stage, we combined expression of the genes into 49 distinct tissue or cell-types. We asked if the exemplar genes for any specific motifs were enriched for specific tissues with a Fisher exact test. After correcting for number of motifs and tissues, we find 251 motifs with genes that are significantly enriched in 23 of the 49 tissue and cell types. For example, the genes associated with the F26A1.1.1 PWM are enriched for pharyngeal genes ($p<5\times10^{-20}$). This PWM is very

similar to the known motif for the TF Pha4 which is known to direct transcription of pharyngeal genes (Gaudet and Mango 2002). In accordance with previous reports that NFI-1 is expressed in muscles (mainly pharynx and head muscles), neurons and intestinal cells (Lazakovitch, Kalb et al. 2005; Lazakovitch, Kalb et al. 2008), our corresponding motif (H01M10.2.1) is also enriched for genes whose GFP-fused promoters are expressed in the pharynx ($p<3x10^{-5}$) and body wall muscle ($p<7x10^{-3}$).

II.4 GO enrichments

Gene Ontology (GO) enrichment has been widely used to assess whether gene sets defined by various clustering methods appear to be significantly related to one another functionally. We compared the exemplar gene sets for each of the PHYLONET PWMs with the GO annotation, at a stringent significance threshold ($p <0.05$ after correction for multiple tests), to find that 3676 (74%) are significantly enriched for at least one biological function

   In C. elegans NFI-1 is shown to be import in regulating motility (Lazakovitch, Kalb et al. 2005). Consistent with this, genes associated with H01M10.2.1 PWM are enriched for GO term microtubule cytoskeleton organization and biogenesis (GO:0000226, 1.0E-06), microtubule organizing center (GO:0005815, 3.7E-06) and microtubule-based process (GO:0007017, 9.3E-06). Vertebrate NF1 is involved in chromatin/chromosome remodeling (Hebbar and Archer 2003) and in vivo target of C. elegans NFI-1 includes many genes involved in this process (Whittle, Lazakovitch et al. 2009). Consistent with this, genes associated with H01M10.2.1 PWM are enriched for GO term centrosome (GO:0005813, 6.8E-07) and spindle organization and biogenesis (GO:0007051, 4.5E-09

). In addition, in vivo NFI-1 targets also includes phosphatase, vaculolar protein sorting factors and protein translocation related proteins and H01M10.2.1 PWM is enriched for GO terms phosphoserine phosphatase activity (GO:0004647, 6.3E-08), vacuolar membrane (GO:0005774, 6.3E-08), vesicle membrane (GO:0012506, 4.3E-06) and protein transport (GO:0015031, 2.9E-06). Taken together, the consistent evidence from multiple independent sources: the similarity of H01M10.2.1 matrix to the C. elegans NFI-1 binding motif and the vertebrate NF-1 binding motif, significant enrichment in tissue-GFP analysis, location analysis, expression analysis as well as significant GO enrichment and NFI-1 targets enrichment, strongly suggests that our PHYLONET-discovered matrix H01M10.2.1 represents the DNA-binding specificity for NFI-1 transcription factor. If we combine all of the biological assays described above we find that a large fraction (4066 of the 4959, 82%) of the predicted motifs have at least one type of evidence to support its regulatory function. Currently, most of the C. elegans TFs are uncharacterized which limits our ability to make direct connections between the PWMs we discover with PHYLONET. But the fact that all of the motifs are conserved across species as well as highly similar in the regulatory regions of multiple genes, and the fact that a large fraction of them are supported by one or more types of experimental or comparative evidence, leads us to believe that they represent regulatory sites for one, or more, TFs and control the expression of C. elegans genes.

III: Using the motifs to predict Cis-regulatory modules (CRMs)

    A cis-regulatory module (CRM) is a segment of DNA that contains multiple transcription factor binding sites which function together to regulate the particular expression patterns of the associated gene. Many studies have shown that in higher organisms CRM is a common strategy in regulating gene expression. If our predicted motifs are functional we would expect the exemplar sites composing those motifs to overlap significantly with experimentally defined regulatory modules. From the literature we collected 41 promoters included in our PHYLONET analysis that have been experimentally tested for the location of regulatory regions. The experiments involve inserting segments of promoters into vectors to create transgenic worms and then it is determined if that region drives expression of a reporter gene, typically GFP. Often the promoter segments that are tested are large and don't provide finer resolution about the critical region, but in other cases the tested segments were small or deletions were introduced to identify critical regions. Using the 2kb upstream sequence of the 41 genes gives us 82kb of potential regulatory sequence for our comparison. There are a total of 61 CRMs that have been experimentally determined in those regions, covering a total of 26,594 bp, 32.4% of the total sequence. This undoubtedly contains regions that are not essential for activity, but that is the limit of the resolution from the currently available experiments. The 41 promoters contain a set of 12,107 exemplar sites and cover 12,473bp, 15.2% of the total sequence. If those two sets of sequences were unrelated we would expect them to overlap by about 5% of the total promoter region, but in fact the overlap is much higher. Of the 61 experimentally confirmed CRMs, 53 (86.9%) of them have overlapping exemplar sites, indicating that using exemplar sites to predict CRMs

would have high sensitivity. 6428 (53.1%) of the exemplar sites are within experimental

CRM regions, which is the minimum positive predictive value (PPV) of the exemplar

sites. It could be much higher because not all regions of the promoters were tested and

there could be additional CRMs in the promoters that are also functional. These results

together indicate that exemplar sites from the PHYLONET analysis can be used to

identify the likely regulatory regions for many *C. elegans* genes.

We can also predict CRMs based on predicted binding sites using the PWMs. While

this will increase the false positive rate, it allows predictions across the whole genome,

not just the ~50% of genes used in the PHYLONET analysis and not limited to the 2 kb

upstream region. We find that the predicted binding sites based on the PWMs are highly

clustered along the promoter sequences (Figure 7), consistent with previous experimental

observations and the general model that DNA sequences with clustered TF binding sites

are usually regulatory sequences that direct specific spatial and temporal gene expression

(Arnone and Davidson 1997; Wasserman and Sandelin 2004; Blanchette, Bataille et al.

2006; Sinha, Liang et al. 2006). To examine whether DNA regions with significantly

enriched motif binding sites correspond to regulatory sequences, we focused on regions

that have binding sites significantly more than average (Z score ≥ 3.09, $p \leq 10^{-3}$). For

example, *hlh-1* (B0304.1) upstream sequence is one of the best studied promoter regions.

A total of six regulatory sequences are identified by detailed deletion and enhancer assays

(Krause, Harrison et al. 1994). The regions with significantly enriched motifs correspond

very well to the experimentally delineated regulatory sequences (Figure 7). Based on this

observation, we developed an algorithm, *C. elegans* Regulatory Module Detector

(CERMOD), to predict regulatory modules using the 4959 PHYLONET PWMs. For *hlh-*

39

1, CERMOD predicted 5 modules in the full 3053 bp upstream sequence which corresponds to all six known regulatory sequences (Figure 7).

To evaluate the predictive power of CERMOD, we performed a thorough literature search to identify any *C. elegans* genes whose promoter regions have been analyzed to locate any regulatory sequences. We identified 75 genes which are expressed in a broad range of tissues at various developmental times (Supplemental Table 1: Experimental Modules). We used upstream intergenic sequences which range from 347 bp to 20,000 bp. There are 149 experimentally determined regulatory regions that are important for corresponding gene expression in neurons, hypoderms, excretory cells, muscle precursor cells, adult muscle cells, vulva cells, sheath cells, etc. These regulatory regions are determined by deletion and/or enhancer assays. Wherever possible, we use regions that are determined by enhancer assay because it better defines the boundary of regulatory regions that are sufficient in regulation. Application of CERMOD on this set of data identified 124 of the 149 (83.2%) experimentally defined modules. Figure 7 shows the comparison between predicted modules with experimentally defined modules in the upstream sequences of 4 well-studied genes. Because some of the predicted modules are located within DNA sequences that have not been tested, we cannot calculate the positive predictive value (PPV) but it is at least 24.3% (214/882). The real PPV is surely higher because in some studies reporter gene expression in tissues other than the interested one is not reported (Wenick and Hobert 2004). Supplemental material: "Pictures of Experimental Module and Predicted Module in Promoter Region" shows the comparison of predicted and experimentally characterized modules in the entire set of genes with experimental evidence.

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 this gene. The simulation is repeated 10,000 times and the sensitivity and PPV are calculated for each one. The average sensitivity is 63.1% with standard deviation of 3.3%. The average PPV is 20.4% with standard deviation of 1.1%. Therefore, the p-values of getting 83.2% sensitivity and 24.3% PPV are both much less than 0.001.

Because many experimental modules have not been further analyzed to delineate the boundary, the functional module can be very long (experimental modules referenced in this manuscript range from 44 to 5287 bp). This resulted in high sensitivity in simulated data. To reduce the effect of those long experimental modules, we used only modules that are within the size range of predicted modules (27 to 580bp) and calculate sensitivity and PPV. The sensitivity did not change much (68 out of 87 are correctly predicted, 78.1%) but the sensitivity of simulated data is greatly reduced (48.5%), making our predictions even more significant.

Cis-regulatory module prediction in miRNA promoters and introns

C. elegans experiments have shown that some introns contain regulatory sequences (Okkema, Harrison et al. 1993; Krause, Harrison et al. 1994; Hwang and Lee 2003). To test if CERMOD can predict CRMs in intron regions we identified 6 genes in which intron regulatory sequence have been mapped in detail. There are 13 experimentally

41

defined modules in the introns from these 6 genes, 10 of which are correctly predicted (76.9%, Figure 8). We performed simulations as described above and the simulated data has an average sensitivity of 46.8%, making our predictions highly significant (p<0.005).

microRNAs (miRNAs) are ~22 nt RNAs that bind to partially imperfectly matched sites on target mRNAs to regulate transcript expression. They are now known to influence a broad range of biological processes. However little is known about how miRNA transcription is regulated. Currently there is only one miRNA, let-7, whose promoter has been dissected to identify regulatory sequences. The let-7 family of microRNAs, first discovered in C. elegans, is functionally conserved from worms to humans. A growing body of evidence suggests that the human let-7 expression is misregulated in many human cancers and restoration of let-7 expression may be a useful therapeutic option in cancers (Boyerinas, Park et al. 2010). Expression of let-7 RNA is temporally regulated with robust expression in the fourth larval and adult states. The DNA fragment located at [-1169, -1285] upstream of the mature RNA is necessary and sufficient for this temporal regulation. We predicted three modules in the ~1.8 kb upstream sequence (Figure 8). The predicted module [-1193, -1259] is completely within the experimentally defined module. let-7 is also expressed in the anchor cell at L3 and in the distal tip cells at the adult stage (Esquela-Kerscher, Johnson et al. 2005). It would be interesting to see whether the other two predicted modules drive let-7 expression in those cells.

Experimental test of CRM prediction

mlc-1 and *mlc-2* are the two muscle regulatory myosin light chain genes in *C. elegans*. They are divergently located and share a 2.6 kb intergenic region. It was shown that they are both expressed in the body-wall muscles, pharyngeal muscles, and vulval muscles. However, the intergenic region has not been analyzed in detail to identify all the regulatory sequences that drive their expression. Previous study has shown that the first 400 bp of *mlc-2* upstream sequence is enough to drive its expression in the body wall muscle cells (GuhaThakurta, Schriefer et al. 2004). To gain better information about transcriptional regulation of *mlc-1* and *mlc-2*, we applied our module prediction method on the intergenic region of *mlc-1/mlc-2* and experimentally tested our prediction. Within this 2662 bp DNA fragment our method predicted 3 CRMs (Figure 9): [39, 203] is just upstream of mlc-1; [1918, 2009] is located at -655 to -746 bp upstream of *mlc-2* translational start codon; [2322, 2489] is close to *mlc-2* translational start codon ATG and corresponds to the first 400 bp upstream that we had previously shown to drive expression in the body wall muscle (GuhaThakurta, Schriefer et al. 2004). We tested regions [1726, 2126], which includes one of the predicted CRMs, and [875, 1747], which does not include any predicted CRM, for enhancer activity by cloning them into a pes-10 minimal promoter (Fire, Harrison et al. 1990). Only the DNA fragment which covers the predicted module showed enhancer activity and the expression was limited to the pharyngeal muscle. So these two experimental results are consistent with the use of PHYLONET PWMs for predicting regulatory regions of *C. elegans* promoters. This result has another interesting aspect. The mlc-2 gene is known to be expressed in both body wall and pharyngeal muscle and we have separated those two tissue specific

43

expression patterns into two separate CRMs. The closest enhancer upsteam of the ATG drives expression only in body wall muscle, and the farther enhancer, located over 500bp upstream, drives expression in the pharyngeal muscle.

IV: Facilitating access to the sites, motifs and module predictions

All data and results discussed here, including the putative regulatory motifs, supporting evidence for each motif, list of motifs that are significant in each analysis, experimental modules and references, pictures of experimental modules and predicted modules in promoter regions as well as in intron regions and microRNA promoters are available via the web interface at http://ural.wustl.edu/~gzhao1/CE_PhyloNet/. Each motif can be accessed by name and the link provides the exemplar sites and the gene list for that motif as well as other related information. Links are also included for all of the genes containing exemplar sites where all of the motifs they are associated with can be found.

We have created files for both the exemplar sites and CERMOD predicted modules across the whole genome in BED formats that can be uploaded as custom tracks and viewed in the UCSC genome browser. These track records can be downloaded from http://ural.wustl.edu/~molee0805/PhyloNet_sites.txt. In the genome browser page, when a Phylonet site or CERMOD module is clicked on it opens up an external site with information about the motif or module. For a motif, this information includes a logo of the motif, the matrix and all exemplar sites genome-wide. Figure 10 highlights the capabilities of this interface.

Conclusions

We performed a genome-wide search for conserved regulatory elements in *C. elegans*, *C. remanei* and *C. briggsae* and identified a total of 4959 regulatory elements. Our study identified regulatory elements with diverse biologically functions which include at least core promoter elements, TF binding sites, and functional RNA sites. Multiple independent evidence provide strong support for their biologically significance. The distribution of these regulatory motifs along promoter sequences is highly clustered which allowed us to accurately detect DNA regulatory sequences that drive spatial/temporal-specific gene expression. Our work greatly expanded our knowledge of regulatory sites in *C. elegans* and is a valuable step towards building a genome-wide regulatory network of *C. elegans*. CERMOD, predicts modules from the distribution of the predicted motif occurrences along the promoter sequences and identifies statistically significantly clustered motif sites. It does not require a training set and it is not necessary to know in which tissue a gene is expressed. it has high sensitivity and specificity on experimentally verified CRMs and we expect it to have similar sensitivity and positive predictive value on any given *C. elegans* sequence. The accessibility of all of our results, the exemplar sites, the predicted motifs and the predicted CRMs, through the UCSC genome browser should make them a valuable resource for the research community.

Acknowledgments

Table 1. Performance of CERMOD on gene promoters.

| Gene | Name | Known Modules | | Predicted Modules | | Overlap | Correctly | Reference |
		Start	End	Start	End	%	predicted	(Krause, Harrison et al. 1994)
B0304.1a	hlh-1	-457	-536	-479	-675	72.5	Yes	
B0304.1a	hlh-1	-725	-949	-724	-816	98.9	Yes	
B0304.1a	hlh-1	-1579	-1932	-1513	-1702	65.3	Yes	
B0304.1a	hlh-1	-2116	-2470	-2353	-2585	50.6	Yes	
B0304.1a	hlh-1	-2537	-2605	-2353	-2585	71	Yes	
B0304.1a	hlh-1	-2633	-2810	-2693	-2771	100	Yes	
B0414.2	rnt-1	-136	-5422	-5263	-5368	100	Yes	(Nam, Jin et al. 2002)
B0414.2	rnt-1	-136	-5422	-3874	-3964	100	Yes	
B0414.2	rnt-1	-136	-5422	-3434	-3499	100	Yes	
B0414.2	rnt-1	-136	-5422	-2795	-2946	100	Yes	
B0414.2	rnt-1	-136	-5422	-1630	-1758	100	Yes	
B0414.2	rnt-1	-136	-5422	-484	-536	100	Yes	
B0414.2	rnt-1	-5874	-6425	-	-	-	No	
B0414.2	rnt-1	-6426	-7150	-6930	-7064	100	Yes	
B0414.2	rnt-1	-6426	-7150	-6231	-6632	51.5	Yes	
C01B7.1b		-1	-667	-291	-476	100	Yes	(Zhao, Schriefer et al. 2007)
C02B8.4	hlh-8	-1	-315	-88	-272	100	Yes	(Harfe, Vaz Gomes et al. 1998)
C02D4.2a	ser-2	-1	-512	-	-	-	No	(Zhao, Schriefer et al. 2007)
C02D4.2c	ser-2	-1	-282	-52	-223	100	Yes	(Wenick and

								Hobert 2004)
C02D4.2c	ser-2	-282	-802	-534	-610	100	Yes	
C02D4.2c	ser-2	-802	-2689	-	-	-	No	
C02D4.2c	ser-2	-2689	-4334	-2984	-3046	100	Yes	
C02D4.2c	ser-2	-2689	-4334	-2826	-2919	100	Yes	
C07H6.7	lin-39	-2000	-5400	-3510	-3691	100	Yes	(Wagmaister, Miley et al. 2006)
C07H6.7	lin-39	-2000	-5400	-1921	-2118	60.1	Yes	
C07H6.7	lin-39	-5100	-6400	-5696	-5796	100	Yes	
C07H6.7	lin-39	-7362	-7700	-	-	-	No	
C08C3.1c	egl-5	-3183	-3486	-3239	-3494	96.9	Yes	(Teng, Girard et al. 2004)
C08C3.1c	egl-5	-5124	-5438	-5375	-5455	79	Yes	
C08C3.1c	egl-5	-5124	-5438	-5213	-5337	100	Yes	
C08C3.1c	egl-5	-7045	-8386	-7777	-7927	100	Yes	
C08C3.1c	egl-5	-7045	-8386	-7249	-7310	100	Yes	
C08C3.1c	egl-5	-7493	-7938	-7777	-7927	100	Yes	
C08C3.1c	egl-5	-7045	-7514	-7249	-7310	100	Yes	
C09D1.1a	unc-89	-1	-588	-365	-557	100	Yes	(GuhaThakurta, Schriefer et al. 2004)
C09D1.1a	unc-89	-1	-588	-136	-289	100	Yes	
C09D1.1a	unc-89	-1	-588	-29	-95	100	Yes	
C10G11.7		-1	-741	-374	-569	100	Yes	(Zhao, Schriefer et al. 2007)
C10G11.7		-1	-741	-81	-180	100	Yes	
C18D1.3	flp-4	-1052	-2385	-1075	-1161	100	Yes	(Etchberger, Lorch et al. 2007)
C33G3.1a	dyc-1	-1	-520	-275	-384	100	Yes	(Zhao, Schriefer et al. 2007)
C33G3.1a	dyc-1	-1	-520	-138	-229	100	Yes	
C36B7.7	hen-1	-1	-1408	-1025	-1117	100	Yes	(Etchberger, Lorch et al. 2007)
C36B7.7	hen-1	-1	-1408	-387	-966	100	Yes	
C36B7.7	hen-1	-1	-1408	-88	-240	100	Yes	
C36B7.7	hen-1	-1	-1408	0	-26	96.3	Yes	
C36B7.7	hen-1	-1556	-1909	-1866	-1928	69.8	Yes	(Wenick and Hobert 2004)
C36B7.7	hen-1	-1556	-1909	-1682	-1769	100	Yes	
C36B7.7	hen-1	-2135	-2911	-2785	-2821	100	Yes	
C36B7.7	hen-1	-2135	-2911	-2615	-2704	100	Yes	
C36B7.7	hen-1	-2135	-2911	-2490	-2582	100	Yes	
C36B7.7	hen-1	-2135	-2911	-2177	-2311	100	Yes	
C36B7.7	hen-1	-2911	-3903	-3168	-3228	100	Yes	
C36B7.7	hen-1	-2911	-3903	-2997	-3057	100	Yes	

C36E6.5	mlc-2	-1	-400	-175	-342	100	Yes	(GuhaThakurta, Schriefer et al. 2004)
C36E6.5	mlc-2	-536	-936	-655	-746	100	Yes	
C37A2.4a	cye-1	-1	-219	-115	-178	100	Yes	(Brodigan, Liu et al. 2003)
C37A2.4a	cye-1	-523	-609	-497	-652	100	Yes	
C37A2.4a	cye-1	-609	-735	-	-	-	No	
C37A2.4a	cye-1	-933	-2200	-	-	-	No	
C37E2.4	ceh-36	-394	-1883	-1032	-1191	100	Yes	(Etchberger, Lorch et al. 2007)
C42D8.2	vit-2	-1	-247	-	-	-	No	(MacMorris, Broverman et al. 1992)
C42D8.8a	apl-1	-6318	-6518	-	-	-	No	(Niwa and Hada 2010)
C54D1.6	bar-1	-1200	-2100	-	-	-	No	(Natarajan, Jackson et al. 2004)
C54D1.6	bar-1	-2000	-3100	-2698	-2795	100	Yes	
C54D1.6	bar-1	-2000	-3100	-2430	-2502	100	Yes	
C54D1.6	bar-1	-2000	-3100	-2277	-2333	100	Yes	
C54D1.6	bar-1	-4779	-5100	-4914	-4997	100	Yes	
C54D1.6	bar-1	-4000	-5100	-4914	-4997	100	Yes	
C54F6.14	ftn-1	-631	-693	-	-	-	No	(Romney, Thacker et al. 2008)
C55B7.12a	che-1	-1	-695	-533	-600	100	Yes	(Etchberger, Lorch et al. 2007)
C55B7.12a	che-1	-1	-695	-88	-240	100	Yes	
D1037.3	ftn-2	-1251	-1313	-1189	-1378	100	Yes	(Romney, Thacker et al. 2008)
E01H11.3	flp-20	-1	-1223	-831	-916	100	Yes	(Etchberger, Lorch et al. 2007)
E01H11.3	flp-20	-1	-1223	-279	-506	100	Yes	
E01H11.3	flp-20	-1	-1223	-70	-143	100	Yes	
E01H11.3	flp-20	-612	-2852	-2402	-2498	100	Yes	
E01H11.3	flp-20	-612	-2852	-2146	-2216	100	Yes	
E01H11.3	flp-20	-612	-2852	-1843	-1933	100	Yes	
E01H11.3	flp-20	-612	-2852	-1679	-1772	100	Yes	
E01H11.3	flp-20	-612	-2852	-831	-916	100	Yes	
EGAP1.3	zmp-1	-2034	-2334	-2194	-2270	100	Yes	(Kirouac and Sternberg 2003)
EGAP1.3	zmp-1	-2034	-2334	-2051	-2134	100	Yes	
F07A5.7	unc-	-1	-500	-90	-401	100	Yes	(GuhaThakur

	15							ta, Schriefer et al. 2004)
F07D3.2	flp-6	-1483	-1950	-	-	-	No	(Etchberger, Lorch et al. 2007)
F08B6.2	gpc-2	-1	-770	-359	-711	100	Yes	(Zhao, Schriefer et al. 2007)
F11C3.3	unc-54	-61	-241	-132	-251	91.7	Yes	(Okkema, Harrison et al. 1993)
F18E9.2	nlp-7	-1052	-2536	-1270	-1420	100	Yes	(Etchberger, Lorch et al. 2007)
F18E9.2	nlp-7	-1052	-2536	-1043	-1224	95.1	Yes	
F22E10.1	pgp-12	-1	-475	-246	-361	100	Yes	(Zhao, Fang et al. 2005)
F27D4.2		-467	-1212	-477	-786	100	Yes	(Zhao, Schriefer et al. 2007)
F29F11.5a	ceh-22	-18	-801	-173	-274	100	Yes	(Kuchenthal, Chen et al. 2001)
F29F11.5a	ceh-22	-18	-801	-14	-133	96.7	Yes	
F29F11.5a	ceh-22	-651	-797	-	-	-	No	
F29F11.5a	ceh-22	-1436	-1922	-	-	-	No	
F31A9.3a	arg-1	-1	-436	-	-	-	No	(Zhao, Wang et al. 2007)
F31A9.3a	arg-1	-1914	-2824	-2505	-2588	100	Yes	
F31A9.3a	arg-1	-1914	-2824	-2155	-2223	100	Yes	
F31A9.3a	arg-1	-2972	-6240	-5452	-5539	100	Yes	
F31A9.3a	arg-1	-2972	-6240	-5212	-5320	100	Yes	
F31A9.3a	arg-1	-2972	-6240	-4562	-4676	100	Yes	
F31A9.3a	arg-1	-2972	-6240	-4278	-4345	100	Yes	
F31A9.3a	arg-1	-2972	-6240	-3666	-3749	100	Yes	
F31A9.3a	arg-1	-2972	-6240	-3300	-3376	100	Yes	
F33D4.3	flp-13	-1	-1036	-850	-959	100	Yes	(Etchberger, Lorch et al. 2007)
F33D4.3	flp-13	-1	-1036	-714	-779	100	Yes	
F33D4.3	flp-13	-1	-1036	-306	-397	100	Yes	
F33D4.3	flp-13	-1	-1036	0	-149	99.3	Yes	
F35D6.1a	fem-1	-1	-170	-28	-218	84.1	Yes	(Gaudet, VanderElst et al. 1996)
F36H1.4a	lin-3	-1	-155	-71	-149	100	Yes	(Hwang and Sternberg 2004)
F38G1.2	egl-	-1	-322	-186	-281	100	Yes	(Cui and Han

	17							2003)
F38G1.2	egl-17	-303	-366	-314	-427	82.8	Yes	
F38G1.2	egl-17	-1409	-1572	-1465	-1587	87.8	Yes	
F38G1.2	egl-17	-2233	-2589	-	-	-	No	
F38G1.2	egl-17	-41	-143	-	-	-	No	(Kirouac and Sternberg 2003)
F38G1.2	egl-17	-234	-392	-314	-427	69.3	Yes	
F38G1.2	egl-17	-234	-392	-186	-281	50	Yes	
F38G1.2	egl-17	-1018	-1526	-1465	-1587	50.4	Yes	
F38G1.2	egl-17	-1097	-1820	-1465	-1587	100	Yes	
F40E10.3	csq-1	-263	-338	-129	-464	100	Yes	(Cho, Eom et al. 1999)
F40E10.3	csq-1	-317	-528	-129	-464	69.8	Yes	
F44F4.13	sra-11	-1189	-2708	-2140	-2229	100	Yes	(Wenick and Hobert 2004)
F44F4.13	sra-11	-1189	-2708	-1992	-2093	100	Yes	
F44F4.13	sra-11	-1189	-2708	-1692	-1775	100	Yes	
F44F4.13	sra-11	-1189	-2708	-1532	-1632	100	Yes	
F44F4.13	sra-11	-1189	-2708	-1357	-1419	100	Yes	
F44F4.13	sra-11	-1189	-2708	-1163	-1300	81.2	Yes	
F44F4.13	sra-11	-2612	-2708	-2637	-2872	74.2	Yes	
F44F4.13	sra-11	-2708	-4823	-4775	-4825	96.1	Yes	
F44F4.13	sra-11	-2708	-4823	-4602	-4737	100	Yes	
F44F4.13	sra-11	-2708	-4823	-4353	-4460	100	Yes	
F44F4.13	sra-11	-2708	-4823	-4188	-4283	100	Yes	
F44F4.13	sra-11	-2708	-4823	-3953	-4036	100	Yes	
F44F4.13	sra-11	-2708	-4823	-3575	-3647	100	Yes	
F44F4.13	sra-11	-2708	-4823	-3464	-3531	100	Yes	
F44F4.13	sra-11	-2708	-4823	-3316	-3386	100	Yes	
F44F4.13	sra-11	-2708	-4823	-2637	-2872	69.9	Yes	
F45D3.2		-1	-697	-434	-520	100	Yes	(Zhao, Schriefer et al. 2007)
F45D3.2		-1	-697	-79	-243	100	Yes	
F46C8.6	dpy-7	-122	-282	-133	-313	93.2	Yes	(Gilleard, Barry et al. 1997)
F48C11.3	nlp-3	-1226	-2488	-2403	-2486	100	Yes	(Etchberger, Lorch et al. 2007)
F48C11.3	nlp-3	-1226	-2488	-2127	-2227	100	Yes	
F48C11.3	nlp-3	-1226	-2488	-1992	-2074	100	Yes	
F48C11.3	nlp-3	-1226	-2488	-1868	-1918	100	Yes	
F48C11.3	nlp-3	-1226	-2488	-1645	-1733	100	Yes	

F52E1.4a	gcy-7	-1	-188	-10	-202	95.2	Yes	(Etchberger, Lorch et al. 2007)
F55B12.1	ceh-24	-1500	-1895	-1545	-1645	100	Yes	(Harfe and Fire 1998)
F55B12.1	ceh-24	-1793	-1910	-	-	-	No	
F55B12.1	ceh-24	-1989	-2443	-	-	-	No	
F55B12.1	ceh-24	-2545	-2602	-	-	-	No	
F55E10.7		-638	-2147	-1192	-1267	100	Yes	(Etchberger, Lorch et al. 2007)
F56D12.5a	vig-1	-191	-596	-408	-607	94.5	Yes	(Shin, Choi et al. 2008)
F58A3.2a	egl-15	-1	-701	-353	-542	100	Yes	(Harfe, Vaz Gomes et al. 1998)
F58A3.2a	egl-15	-1	-701	-72	-301	100	Yes	
F58B4.1a	nas-31	-1	-312	-24	-228	100	Yes	(Zhao, Fang et al. 2005)
H14A12.4	mls-1	-1	-798	-115	-322	100	Yes	(Zhao, Wang et al. 2007)
K03D10.1	kal-1	-1	-270	-73	-119	100	Yes	(Wenick and Hobert 2004)
K03D10.1	kal-1	-1548	-2865	-2828	-2886	64.4	Yes	
K03D10.1	kal-1	-1548	-2865	-2635	-2714	100	Yes	
K03D10.1	kal-1	-2865	-3679	-	-	-	No	
K03D10.1	kal-1	-3679	-5256	-4760	-4836	100	Yes	
K03D10.1	kal-1	-3679	-5256	-4150	-4224	100	Yes	
K03D10.1	kal-1	-3679	-5256	-3929	-4021	100	Yes	
K03E6.1	lim-6	-1	-208	-	-	-	No	(Etchberger, Lorch et al. 2007)
K07C6.4	cyp-35B1	-1	-360	-107	-284	100	Yes	(Iser, Wilson et al. 2011)
K07C6.4	cyp-35B1	-510	-660	-351	-632	81.5	Yes	
K10G6.3		-378	-929	-746	-859	100	Yes	(Zhao, Schriefer et al. 2007)
K10G6.3		-378	-929	-606	-693	100	Yes	
K10G6.3		-378	-929	-443	-540	100	Yes	
K12F2.1	myo-3	-328	-749	-220	-639	74.3	Yes	(Okkema, Harrison et al. 1993)
K12F2.1	myo-3	-1010	-1948	-1441	-1538	100	Yes	
K12F2.1	myo-3	-1010	-1948	-1246	-1335	100	Yes	

R02E12.6	hrg-1	-1	-254	-193	-282	68.9	Yes	
R03C1.3a	cog-1	-2924	-3444	-2837	-3061	61.3	Yes	(Etchberger, Lorch et al. 2007)
R06C7.10	myo-1	-123	-500	-156	-479	100	Yes	(Okkema, Harrison et al. 1993)
R06C7.10	myo-1	-646	-1725	-1416	-1713	100	Yes	
R13A5.5	ceh-13	-2334	-2604	-2383	-2461	100	Yes	(Stoyanov, Fleischmann et al. 2003)
R13A5.5	ceh-13	-2000	-3200	-2851	-2964	100	Yes	(Streit, Kohler et al. 2002)
R13A5.5	ceh-13	-2000	-3200	-2661	-2743	100	Yes	
R13A5.5	ceh-13	-2000	-3200	-2383	-2461	100	Yes	
R13A5.5	ceh-13	-3200	-3940	-3798	-3886	100	Yes	
R13A5.5	ceh-13	-3200	-3940	-3397	-3486	100	Yes	
R13A5.5	ceh-13	-6150	-6600	-6220	-6284	100	Yes	
T01E8.2	ref-1	-282	-435	-	-	-	No	(Neves, English et al. 2007)
T03D8.5	gcy-22	-1	-829	-377	-612	100	Yes	(Etchberger, Lorch et al. 2007)
T03D8.5	gcy-22	-1	-829	0	-168	99.4	Yes	
T15H9.3	hlh-6	-1	-241	-28	-117	100	Yes	(Raharjo and Gaudet 2007)
T15H9.3	hlh-6	-241	-747	-502	-621	100	Yes	
T18D3.4	myo-2	-370	-686	-542	-630	100	Yes	(Okkema, Harrison et al. 1993)
T18D3.4	myo-2	-458	-764	-542	-630	100	Yes	
T18D3.4	myo-2	-17	-239	-79	-153	100	Yes	
T20G5.6	unc-47	-1	-239	-54	-362	77.8	Yes	(Eastman, Horvitz et al. 1999)
T22B7.1a	egl-13	-1	-1330	-1247	-1330	100	Yes	(Oommen and Newman 2007)
T22B7.1a	egl-13	-1	-1330	-521	-605	100	Yes	
T22B7.1a	egl-	-1	-1330	-213	-294	100	Yes	

	13							
T22B7.1a	egl-13	-1	-1330	-73	-136	100	Yes	
T22B7.1a	egl-13	-1	-632	-521	-605	100	Yes	
T22B7.1a	egl-13	-1	-632	-213	-294	100	Yes	
T22B7.1a	egl-13	-1	-632	-73	-136	100	Yes	
T28B8.1		-1	-597	-187	-546	100	Yes	(Zhao, Schriefer et al. 2007)
T28B8.1		-1	-597	0	-110	99.1	Yes	
W03A3.1	ceh-10	-1	-1121	-458	-545	100	Yes	(Wenick and Hobert 2004)
W03A3.1	ceh-10	-1	-1121	-16	-324	100	Yes	
W03A3.1	ceh-10	-2851	-3655	-2976	-3324	100	Yes	
W06H8.6		-1	-675	-469	-820	58.8	Yes	(Zhao, Schriefer et al. 2007)
W06H8.6		-675	-1333	-858	-1149	100	Yes	
W09B12.1	ace-1	-1	-288	0	-173	99.4	Yes	(Culetto, Combes et al. 1999)
W09B12.1	ace-1	-493	-698	-610	-706	91.8	Yes	
W09B12.1	ace-1	-1731	-2049	-1953	-2093	68.8	Yes	
W09B12.1	ace-1	-1731	-2049	-1787	-1865	100	Yes	
W09B12.1	ace-1	-2049	-2398	-2277	-2374	100	Yes	
Y105E8B.1a	tmy-1	-1	-818	-78	-389	100	Yes	(Kagawa, Sugimoto et al. 1995)
Y105E8B.1c	tmy-1	-1	-853	-327	-455	100	Yes	(Anyanful, Sakube et al. 2001)
Y22F5A.3	snap-25	-980	-1123	-	-	-	No	(Hwang and Lee 2003)
Y22F5A.3	snap-25	-169	-305	-	-	-	No	
Y37D8A.23a	unc-25	-1	-180	-10	-92	100	Yes	(Eastman, Horvitz et al. 1999)
Y73C8B.4	lag-2	-1	-1029	-279	-400	100	Yes	
Y73C8B.4	lag-2	-5567	-5674	-5628	-5714	54	Yes	
ZC247.3	lin-11	-1	-1700	-1288	-1355	100	Yes	(Gupta and Sternberg 2002)
ZC247.3	lin-11	-1	-1700	0	-245	99.6	Yes	
ZC247.3	lin-11	-1700	-2230	-1758	-1818	100	Yes	
ZC247.3	lin-11	-2230	-2880	-2722	-2833	100	Yes	
ZC247.3	lin-11	-2230	-2880	-2339	-2435	100	Yes	

ZC416.8a	unc-17	-1	-822	-113	-289	100	Yes	(Wenick and Hobert 2004)
ZC416.8a	unc-17	-822	-1495	-	-	-	No	
ZC416.8a	unc-17	-1876	-2463	-	-	-	No	
ZK112.7	cdh-3	-1044	-1607	-1030	-1324	95.3	Yes	(Kirouac and Sternberg 2003)
ZK112.7	cdh-3	-2877	-3629	-2869	-3093	96.4	Yes	
ZK112.7	cdh-3	-3519	-3751	-3592	-3690	100	Yes	
ZK455.7	pgp-3	-1	-502	-245	-363	100	Yes	(Zhao, Fang et al. 2005)
ZK455.7	pgp-3	-1	-502	-120	-203	100	Yes	
ZK455.7	pgp-3	-1	-502	0	-62	98.4	Yes	
ZK652.5	ceh-23	-1	-791	-546	-667	100	Yes	(Wenick and Hobert 2004)
ZK652.5	ceh-23	-1	-791	-224	-314	100	Yes	
ZK652.5	ceh-23	-791	-1109	-863	-938	100	Yes	
ZK652.5	ceh-23	-2269	-2453	-2376	-2467	84.8	Yes	
ZK652.5	ceh-23	-4915	-5455	-5086	-5315	100	Yes	
ZK652.5	ceh-23	-4915	-5455	-4878	-4971	60.6	Yes	
ZK652.5	ceh-23	-5486	-6376	-6012	-6099	100	Yes	
ZK652.5	ceh-23	-5486	-6376	-5817	-5878	100	Yes	
ZK652.5	ceh-23	-5486	-6376	-5669	-5774	100	Yes	
ZK652.5	ceh-23	-5486	-6376	-5475	-5544	84.3	Yes	
ZK652.5	ceh-23	-6776	-6819	-6688	-6822	100	Yes	
ZK652.5	ceh-23	-6819	-7275	-7100	-7238	100	Yes	
ZK970.6	gcy-5	-1	-306	-71	-161	100	Yes	(Wenick and Hobert 2004)

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.

Brodigan, T.M., Liu, J., Park, M., Kipreos, E.T., and Krause, M. (2003). Cyclin E expression during development in Caenorhabditis elegans. Dev Biol *254*, 102-115.

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.

Cui, M., and Han, M. (2003). Cis regulatory requirements for vulval cell-specific expression of the Caenorhabditis elegans fibroblast growth factor gene egl-17. Dev Biol *257*, 104-116.

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.

Eastman, C., Horvitz, H.R., and Jin, Y. (1999). Coordinated transcriptional regulation of the unc-25 glutamic acid decarboxylase and the unc-47 GABA vesicular transporter by the Caenorhabditis elegans UNC-30 homeodomain protein. J Neurosci *19*, 6225-6234.

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.

Gaudet, J., VanderElst, I., and Spence, A.M. (1996). Post-transcriptional regulation of sex determination in Caenorhabditis elegans: widespread expression of the sex-determining gene fem-1 in both sexes. Mol Biol Cell *7*, 1107-1121.

Gilleard, J.S., Barry, J.D., and Johnstone, I.L. (1997). cis regulatory requirements for hypodermal cell-specific expression of the Caenorhabditis elegans cuticle collagen gene dpy-7. Mol Cell Biol *17*, 2301-2311.

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.

Gupta, B.P., and Sternberg, P.W. (2002). Tissue-specific regulation of the LIM homeobox gene lin-11 during development of the Caenorhabditis elegans egg-laying system. Dev Biol *247*, 102-115.

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. (1998). Analysis of a Caenorhabditis elegans Twist homolog identifies conserved and divergent aspects of mesodermal patterning. Genes Dev *12*, 2623-2635.

Hwang, B.J., and Sternberg, P.W. (2004). A cell-specific enhancer that specifies lin-3 expression in the C. elegans anchor cell for vulval development. Development *131*, 143-151.

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.

Iser, W.B., Wilson, M.A., Wood, W.H., 3rd, Becker, K., and Wolkow, C.A. (2011). Co-regulation of the DAF-16 target gene, cyp-35B1/dod-13, by HSF-1 in C. elegans dauer larvae and daf-2 insulin pathway mutants. PLoS One *6*, e17369.
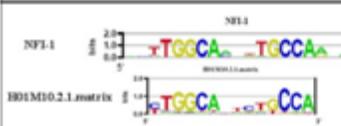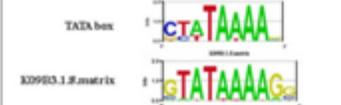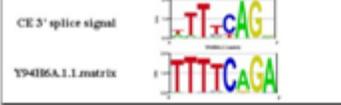
Kagawa, H., Sugimoto, K., Matsumoto, H., Inoue, T., Imadzu, H., Takuwa, K., and

> Sakube, Y. (1995). Genome structure, mapping and expression of the
>
> tropomyosin gene tmy-1 of Caenorhabditis elegans. J Mol Biol *251*, 603-613.

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.

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.

Kuchenthal, C.A., Chen, W., and Okkema, P.G. (2001). Multiple enhancers contribute to

> expression of the NK-2 homeobox gene ceh-22 in C. elegans pharyngeal muscle.
>
> Genesis *31*, 156-166.

MacMorris, M., Broverman, S., Greenspoon, S., Lea, K., Madej, C., Blumenthal, T., and

> Spieth, J. (1992). Regulation of vitellogenin gene expression in transgenic
>
> Caenorhabditis elegans: short sequences required for activation of the vit-2
>
> promoter. Mol Cell Biol *12*, 1652-1662.

Nam, S., Jin, Y.H., Li, Q.L., Lee, K.Y., Jeong, G.B., Ito, Y., Lee, J., and Bae, S.C.

> (2002). Expression pattern, regulation, and biological role of runt domain
>
> transcription factor, run, in Caenorhabditis elegans. Mol Cell Biol *22*, 547-554.

Natarajan, L., Jackson, B.M., Szyleyko, E., and Eisenmann, D.M. (2004). Identification

> of evolutionarily conserved promoter elements and amino acids required for
>
> function of the C. elegans beta-catenin homolog BAR-1. Dev Biol *272*, 536-557.

Neves, A., English, K., and Priess, J.R. (2007). Notch-GATA synergy promotes

    endoderm-specific expression of ref-1 in C. elegans. Development *134*, 4459-

    4468.

Niwa, R., and Hada, K. (2010). Identification of a spatio-temporal enhancer element for

    the Alzheimer's amyloid precursor protein-like-1 gene in the nematode

    Caenorhabditis elegans. Biosci Biotechnol Biochem *74*, 2497-2500.

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.

Oommen, K.S., and Newman, A.P. (2007). Co-regulation by Notch and Fos is required

    for cell fate specification of intermediate precursors during C. elegans uterine

    development. Development *134*, 3999-4009.

Raharjo, I., and Gaudet, J. (2007). Gland-specific expression of C. elegans hlh-6 requires

    the combinatorial action of three distinct promoter elements. Dev Biol *302*, 295-

    308.

Romney, S.J., Thacker, C., and Leibold, E.A. (2008). An iron enhancer element in the

    FTN-1 gene directs iron-dependent expression in Caenorhabditis elegans

    intestine. J Biol Chem *283*, 716-725.

Shin, K.H., Choi, B., Park, Y.S., and Cho, N.J. (2008). Analysis of C. elegans VIG-1

    expression. Mol Cells *26*, 554-557.

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.

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.

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.

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.

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.

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.

Figure 6

Examples of three different classes of conserved elements and supporting evidence.
Upper panel: Sequence logo, reference, binding factor as well as supporting evidence for
NFI-1, TATA box and C. elegans 3' splice/trans-splicing signal. Lower panel:
Distribution of K09B3.1.8.matrix and Y94H6A.1.1.matrix sites on promoters.

Figure 7

Comparison between predicted Cis-Regulatory Module (CRM) with experimentally defined CRM in four best studied promoters. A. Comparison between predicted CRM with experimentally defined CRM in hlh-1. B. Distribution of Z score of number of motif sites across the hlh-1 promoter. C. Comparison between predicted Cis-Regulatory Module (CRM) with experimentally defined CRM in myo-3, myo-1 and myo-2. Turquoise bar: experimentally tested DNA fragment without regulatory function; Red bar: experimentally tested DNA fragment with regulatory function; Deep blue bar: promoter sequence; Grey bar: predicted CRM. Black triangle: translational start codon. Position coordinates shown are relative to translational start codon.

Figure 8

Comparison between predicted CRM with experimentally defined CRM in intron regions and in miRNA let-7 promoter. A. Comparison in intron regions. B. Comparison in miRNA let-7 promoter. Red bar: experimentally tested DNA fragment with regulatory function; Deep blue bar: input DNA sequence; Grey bar: predicted CRM. Black triangle: translational start codon. Position coordinates shown are relative to translational start codon.
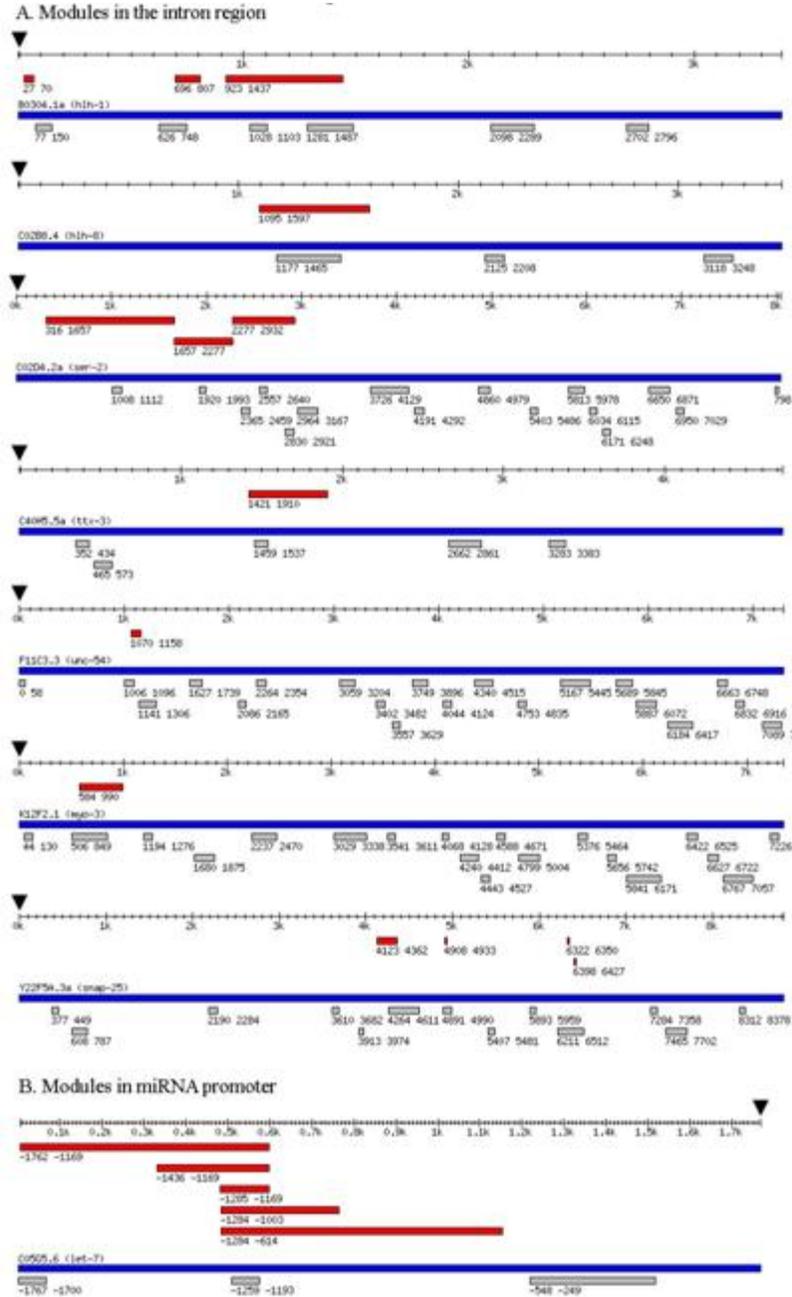
Figure 9

Experimental test of predicted CRM in mlc-1/mlc-2 intergenic region. Turquoise bar: DNA fragment tested that did not show regulatory function; Deep blue bar: mlc-1/mlc-2 intergenic region DNA sequence; Grey bar: predicted CRM. Black triangle: translational start codon. Positive po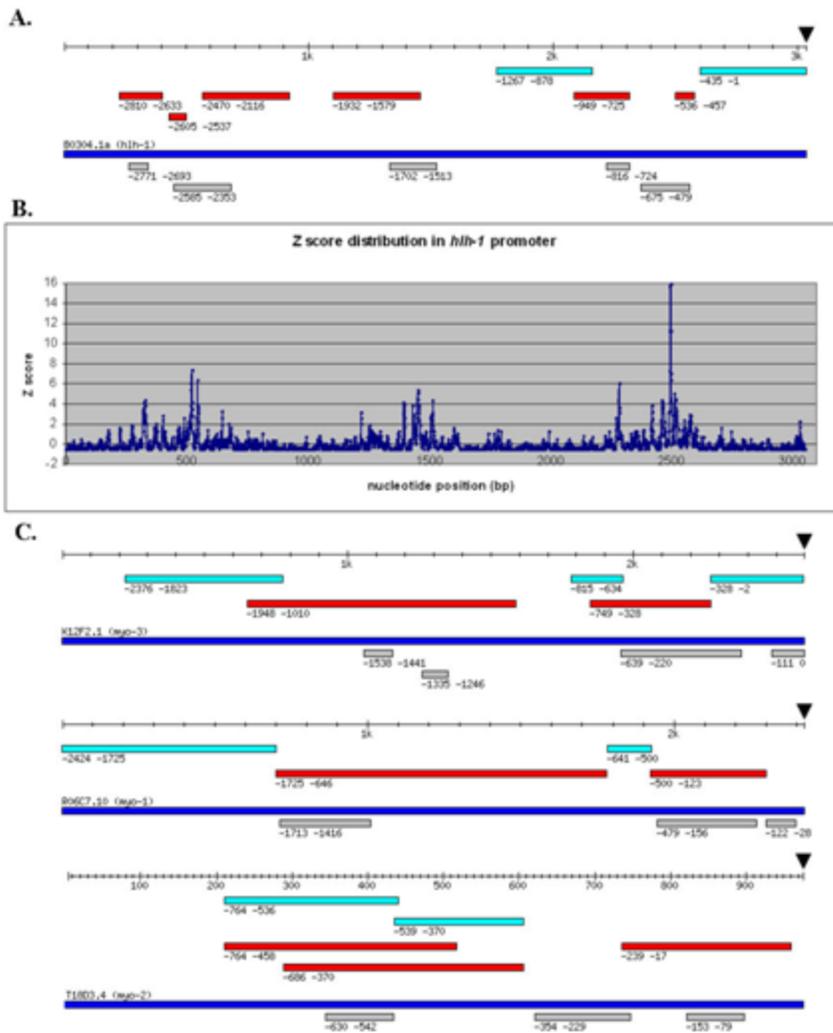sition coordinates shown are relative to mlc-1 translational start codon. Negative position coordinates shown are relative to mlc-2 translational start codon.
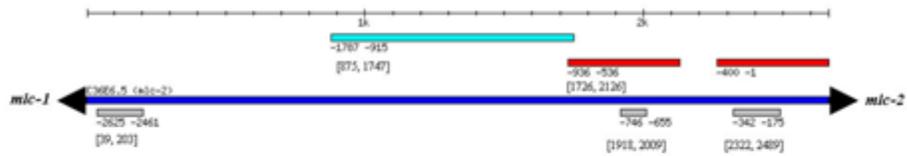
Figure 10

Example use of the UCSC genome browser. A. Screen shot of genomic region containing exemplar sites; clicking on the red circled exemplar site in front of the gene F26A3.6 takes you to the additional information page for this gene, shown in part B. Clicking on the outside link (highlighted in red) takes you to a table with all the motifs in this promoter region, shown in part C. Clicking on the specific motif highlighted in red opens a new page displaying the additional information for this motif, shown in part D.

Figure 10: A
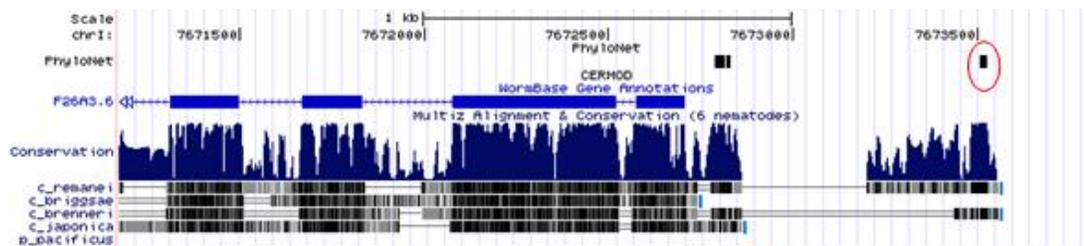
Figure 10: B
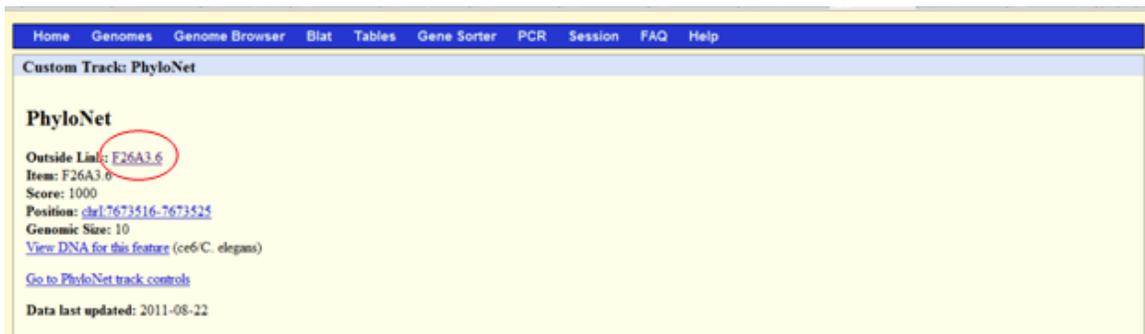
Figure 10: C
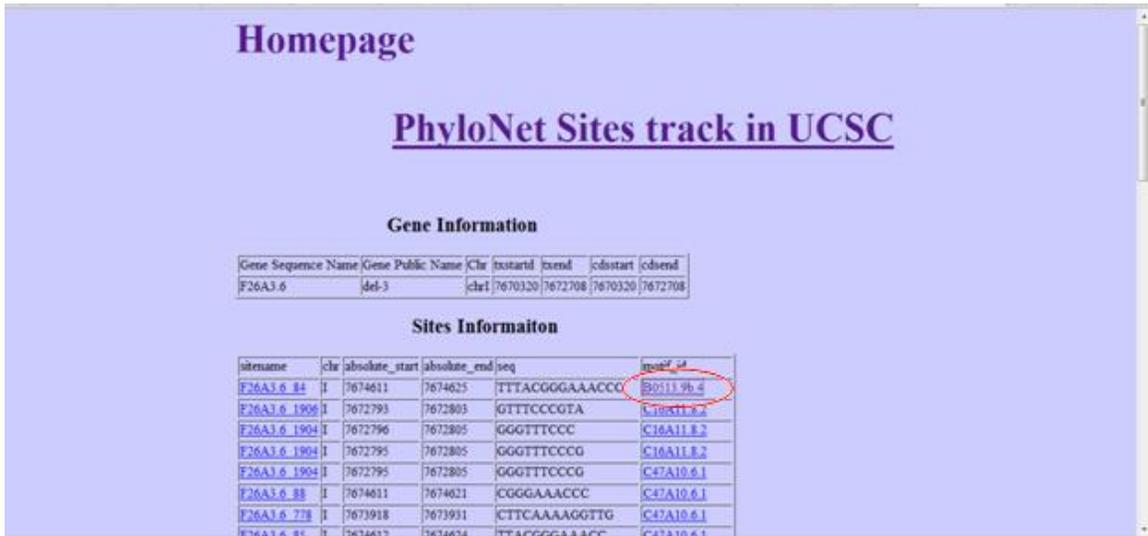
Figure 10: D
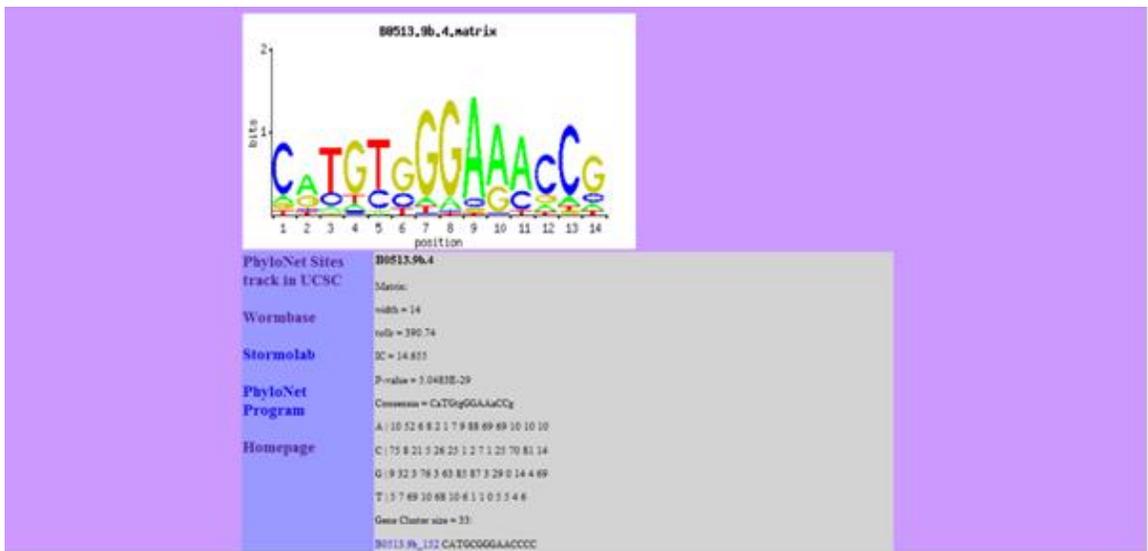
Supplemental Figure 1

Comparison between discovered motifs and previously characterized transcription factor motifs.
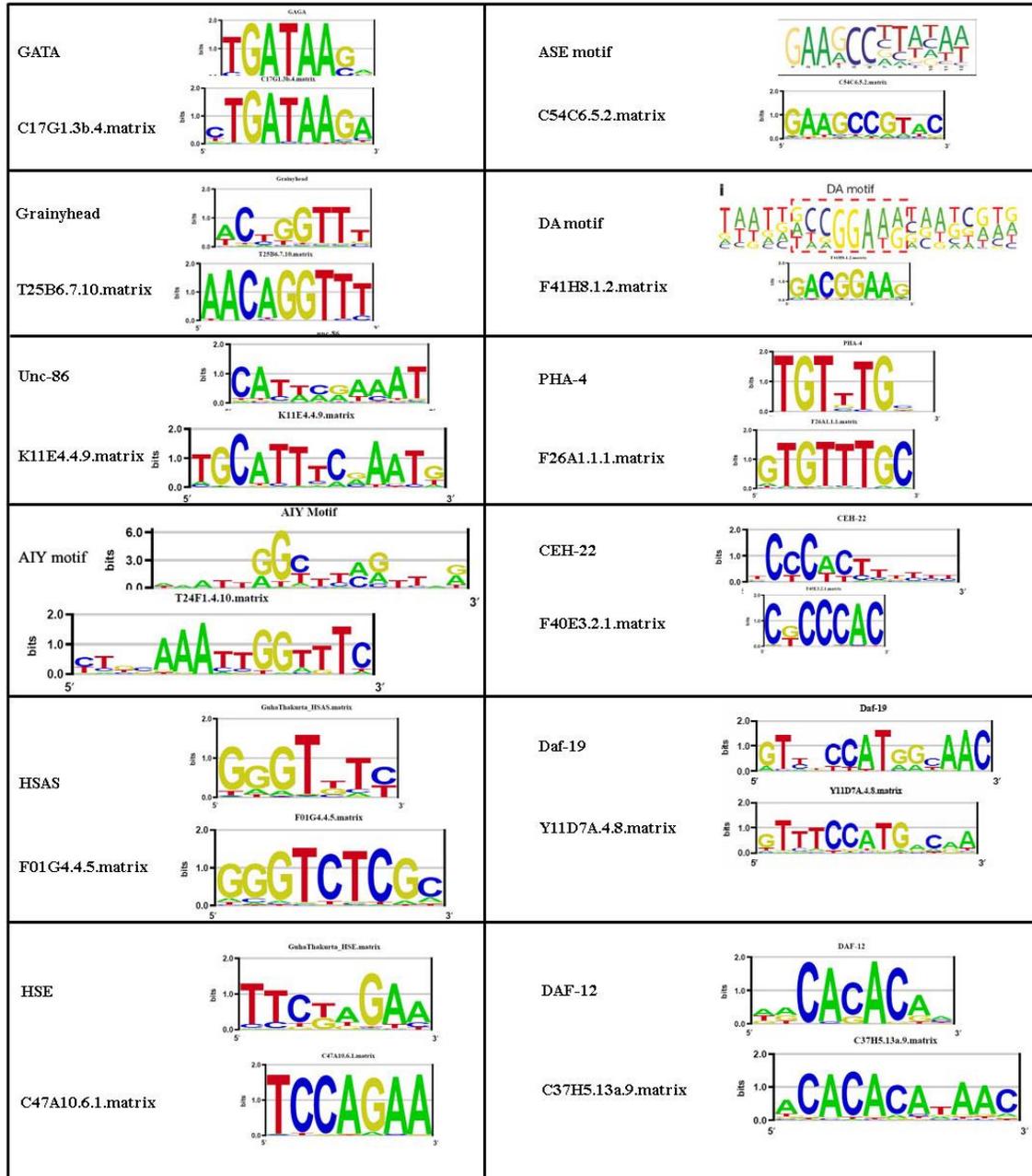
Supplemental Figure 2

Comparison between discovered motifs and previously characterized regulatory elements.
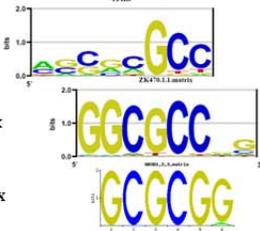
Chapter 3: Discovering conserved *cis*-Regulatory elements in C. elegans using Magma[2]

Abstract

Regulatory sites that control gene expression are essential to the proper functioning of cells, and identifying them is critical for modeling regulatory networks. We have developed Magma (Multiple Aligner of Genomic Multiple Alignments), a software tool for multiple species, multiple gene motif discovery. Magma identifies putative regulatory sites that are conserved across multiple species and occur near multiple genes throughout a reference genome. It is particularly designed to be very efficient in the discovery of regulatory elements from higher-order eukaryotes with large non-coding regions. Magma takes as input multiple alignments from non-coding regions, which can include gaps. It computes similarities between profiles of conserved non-coding regions, clusters similar profiles into motifs and reduces any resulting redundancy in an efficient manner. Magma is about 70 times faster than PhyloNet, a previous program for this task, with slightly greater sensitivity. We ran Magma on all non-coding DNA conserved between C. elegans and 5 additional species, about 80Mbp in total, in less than 4 hours. We obtained 2309 motifs with lengths from 6-20bp, each occurring at least 10 times throughout the genome, that collectively covered about 500kbp of the genomes, approximately 0.6% of the input. Predicted sites occurred in all types of non-coding sequence but were especially enriched in the promoter regions. Comparisons to several experimental datasets show that Magma motifs correspond to a variety of known regulatory motifs. Finally we show that Magma has tractable scaling with reasonable runtimes for other high-order eukaryotes.

Introduction

A key area of genomic research is understanding the *cis*-regulatory network that governs

transcriptional regulation. Over the past two decades, many computational approaches

have been developed to discover transcription factor (TF) binding sites in the genome by

identifying recurring sequence motifs that bind a particular factor. Discovering such

motifs is challenging because they are usually short (5-12 bases) and degenerate.

Traditional algorithms to recognize motifs in genomic DNA take one of two basic

approaches. The *multiple gene, single species* approach recognizes motifs because they

recur with few changes in the promoters of multiple genes within a single genome

(Lawrence, Altschul et al. 1993; Hertz and Stormo 1999; Bailey, Williams et al. 2006;

Elemento, Slonim et al. 2007). In contrast, the *single gene, multiple species* – or

phylogenetic footprinting – approach recognizes motifs in a single promoter region by

their conservation across species, which is assumed to be greater than that of the

surrounding background sequence (Gelfand 1999; McGuire, Hughes et al. 2000; McCue,

Thompson et al. 2001; Panina, Mironov et al. 2001; Rajewsky, Socci et al. 2002; Frazer,

Elnitski et al. 2003; Panina, Vitreschak et al. 2003; Marchal, De Keersmaecker et al.

2004). These methods work because binding sites are typically under selective pressure

and therefore mutate more slowly than the surrounding sequence. Wang and Stormo

(2003) combined these two approaches in their PhyloCon program, which uses

alignments of orthologous promoter regions rather than individual DNA sequences. In

this paradigm, a motif is required both to recur across different promoters and to be

conserved across species in each of its occurrences. Other tools that take a conceptually

similar approach include (Qin, McCue et al. 2003; Jensen, Shen et al. 2005; Monsieurs,

Thijs et al. 2006), all of which report results on bacterial promoters.

To scale PhyloCon's methods to discover motifs across an entire genome, the successor program—PhyloNet (Wang and Stormo 2005)—implemented a BLAST-like seeded alignment algorithm to accelerate detection of putative motif instances across thousands of promoters. This allowed its application to all noncoding sequences of the yeast genome, but still at a high cost – over five CPU-days on a 2.4GHz workstation. The noncoding sequences of a higher eukaryotic genome represent tens to hundreds of times more sequence than yeast. Most phylogenetically-based motif-finding algorithms scale quadratically with the input size, so the lengthy times expected for higher eukaryotic promoter analyses are a deterrent to genome-wide motif discovery.

This work describes Magma (Multiple Aligner of Genomic Multiple Alignments), a new algorithm for multi-gene, multi-species computational motif discovery. Magma significantly departs from the PhyloNet pipeline for accelerated operations, most substantially by introducing new algorithms to group putative TF binding sites into motifs and to reduce redundancy in its output. Magma also operates on gapped genomic sequence alignments. Using alignments of *Saccharomyces* promoters, Magma runs almost 70 times faster than PhyloNet with improved sensitivity. Magma scales to analyses of higher eukaryotes; it can analyze all proximal promoters in *Drosophila* in less time than that required by PhyloNet to analyze yeast. Although Magma's efficiency allows us to perform whole-genome motif-finding on higher eukaryotes, its motif-finding methods can sometimes produce many redundant, partially overlapping motifs. We alleviate this problem by with a fast, greedy, set-covering approach (Chvatal 1979).

We demonstrate Magma's motif discovery prowess using essentially all of *C. elegans* non-coding sequence: an 70Mbp-search space consisting of promoters, UTRs, introns, and downstream regions. To the best of our knowledge, this is the most comprehensive motif-finding effort to date in *C. elegans*. Furthermore, we show that these motifs and their conserved exemplar sites correspond to many known regulatory sites, are enriched in TF-bound regions, and are correlated with expression. Magma and all post-processing software are available for noncommercial use by request to the authors.

Methods

The Magma Computation

Magma takes as input a collection of multiple sequence alignments or *profiles* (e.g. the Multiple Alignment Format, or MAF, blocks from UCSC), each of which aligns orthologous genomic sequences from different species. Its goal is to discover short *motifs*, which are approximate sequence patterns that occur in multiple instances, or *exemplar sites*, within each genome and appear distinct from the surrounding sequence. However, because Magma searches profiles rather than single sequences, each instance of a motif is itself a collection of aligned sequences exhibiting significant conservation across the species in its profile. Magma compares pairs of profiles using the *average log-likelihood ratio (ALLR) score*, a measure of similarity between columns of two multiple alignments (Wang and Stormo 2003). The ALLR is well-defined for pairs of columns containing different total numbers of characters, so it may be applied to columns which have different number of bases due to gaps. For two motifs of equal length, their total ALLR score is simply the sum of the ALLR scores of their corresponding columns, ignoring gapped positions.

Magma discovers motifs by comparing one input profile, the *query*, to a database of all other profiles. Each profile in the input serves as the query in turn, until all profiles have been compared pairwise. Magma's search has two phases: generation of *high-scoring segment pairs (HSPs)*, which locally align two profiles, and clustering of all HSPs involving a given query to form motifs. HSP generation is further subdivided into seed matching and extension.

An HSP is a local alignment of the query profile and a database profile, such that the total ALLR score of all aligned column pairs exceeds a user-defined threshold T. To reduce the computational cost of search, and to allow identification of multiple HSPs per profile pair, HSP generation uses a *seeded alignment* approach on a simplified representation of the input profiles. Each input profile is first quantized into a sequence over an alphabet of 15 symbols, each of which represents a particular vector of base counts, by mapping each profile column to the symbol whose vector has the most similar distribution (Wang and Stormo 2005). The alignment score for a pair of symbols is the ALLR score for the corresponding pair of vectors. The quantized query and database profiles are scanned for *seed matches*, or pairs of fixed-length substrings with at least some minimum score, using a neighborhood hashing strategy analogous to that used by BLASTP for sequence alignment. Each seed match between two profiles is extended by dynamic programming into the best HSP passing through the match, and HSPs with scores exceeding T are retained. Whereas seed matching is done on the quantized profiles, extension is done in the original profiles using the full ALLR score.

Magma's Clustering Algorithm

The clustering phase collects and aligns putative motif instances from the HSPs generated by the previous phase. A c*luster* is a collection of HSPs, all of which overlap on a given query profile P_q. A cluster of n HSPs therefore defines intervals from at most n distinct profiles besides the query, all of which are aligned to P_q (and hence transitively to each other). Clustering first groups all HSPs for a query, then reduces each cluster to a single motif, with each interval possibly contributing one motif instance. A motif may use only a subset of the cluster's intervals, and each interval must be adjusted so that all instances

of the motif have the same length. Subsetting and length adjustment are performed so as to maximize the sum of ALLR scores between the instance drawn from P_q and each other instance in the motif.

Magma uses efficient clustering methods that offer strong performance and quality guarantees. Edges of an HSP overlap graph are determined by overlaps between intervals on the same profile, making this graph an *interval graph*. All maximal cliques in such a graph can easily be found in time linear in the number of HSPs and enumerated in time proportional to their total size (Gupta, Lee et al. 1982). Magma therefore uses interval clique finding to guarantee both maximality and exhaustive enumeration of clusters, with much better scalability than general clique finding. To avoid building clusters from HSPs that overlap by very little (e.g. a single base), it is desirable to enforce a minimum overlap of *k* positions to create an edge in the overlap graph. Magma enforces this criterion by reducing each interval's right endpoint by *k-1* positions prior to clique finding.

To simplify conversion of clusters to motifs, Magma uses the following enumerative algorithm. For each HSP H_j in the cluster, let P_q (the query) and P_j be the profiles that it aligns, and let $[l_j, r_j]$ and $[l'_j, r'_j]$ be the intervals that it aligns from P_q and P_j, respectively. Let $d_j = l'_j - l_j$ be the *diagonal* of H_j, that is, the offset of its starting indices in the query and database profiles.

Suppose that the HSPs in a cluster have $min_j l_j = L$ and $max_j r_j = R$. For each left endpoint ℓ and right endpoint r, $L \leq \ell \leq r \leq R$, we find the best-scoring motif whose instance on P_q is the interval $[\ell, r]$. The instance corresponding to HSP H_j is then $[\ell + d_j, r + d_j]$. (If this instance runs off either end of P_j, then it is discarded for this choice of

endpoints.) We then discard any instance whose ALLR score versus the query instance is negative and retain the total score $s_{\ell,r}$ of the remaining instances. The motif with the highest total ALLR score for the cluster is the one with endpoints $\text{argmax}_{\ell,r}\, s_{\ell r}$ in profile P_q.

Our enumerative algorithm requires time $\Theta(m^2 n)$, where n is the number of HSPs in the cluster and $m = R - L + 1$. However, the ALLR scores for each column of the alignment between each P_j and P_q can be precomputed and stored in total time $\Theta(mn)$. Hence, the constant factor associated with the quadratic cost in m is small in practice, consisting mostly of addition and table lookup. We also note that when the goal is instead to minimize the statistical p-value defined in (Wang and Stormo 2005) for the motif, the motif with best p-value for a cluster can still be found in time $\Theta(m^2\, n \log n)$.

Reducing Redundant Motifs

The motifs obtained by HSP finding and clustering may contain many overlapping, partially redundant motifs. The major source of redundancy is the re-use of overlapping profiles in construction of multiple motifs. Since we know the genomic coordinates of all the exemplar sites that were used to construct every motif, we can re-describe this problem as an NP-Complete Set-Covering problem (Karp 1972; Vazirani 2001). Given a universe U of exemplar contigs (i.e. contiguous regions built from overlapping exemplar sites) and a collection of motifs S, each of which covers a subset of U, a cover is a subset C of S whose union of exemplar sites covers all of U.

We implement a fast greedy approximation for the Set-Covering problem to significantly reduce the motif redundancy in the final output. Greedy algorithms for minimum Set-

Covering achieve a log(n) approximation, where n is the size of the largest set (Chvatal 1979):

$$H(n) = \sum_{k=1}^{n} \frac{1}{k} \cong \ln(n)$$

This means we use at most log(n) times the minimum number of motifs needed to cover all instances. Our implementation is similar to other Set-Covering solutions but with some slight modifications. At each iteration, we define a cover as the set of sites from the most occurring motif (m*), as well as sites from any other motif that overlaps m* sites by at least *d* sites. Thus at each iteration we remove a set of sites u* in U and their associated motifs from the problem. We continue this recursion as long as $|u^*| \geq M_u$ minimum unique sites. The redundant motifs in each resulting cover are subsequently resolved by iteratively scanning all the sites with each motif (by order of most occurrences) and masking their instances. This continues until there are fewer than M_u sites left in the cover.

Results

Magma is a fast genome-wide motif-finder with tractable scaling for higher-order eukaryotes

Magma was designed in part to overcome performance limitations in the earlier PhyloNet motif-finding software. To measure Magma's performance relative to PhyloNet, we ran both programs to discover initial motifs in yeast promoters. On a cluster of 2.4GHz AMD Opteron processors, we observed a ~70x speedup. Moreover, Magma's ability to use gapped profiles, which better aligns motif instances in different parts of the same profile, allowed it to discover more known motifs than PhyloNet while still including less of the reference sequence in its output. We also examined how Magma scales when applied to more complex eukaryotes (Table 2).

Table 2: Magma scales to higher-order eukaryotes with practical runtime

Organism	Search Space (Mbp)	Magma-DiscoveryTime (cpu secs)
S. cereviasae	1.74	101
D. melanogaster	15.36	3184
C. elegans	69.10	12915

Running Magma on *D. melanogaster*'s conserved promoter regions (~9x increase in search space) required about 30x more time than the yeast experiment. The complete *C. elegans* conserved regulome from six species (~40x search space) required ~130x more time (~3.5hrs). In practice, we implement Magma such that the set of all queries is

distributed across several processors, so that the actual running time for *C. elegans* was only ~0.75 hrs.

Characteristics of Magma *C. elegans* Motifs

We discovered 2,309 motifs in *C. elegans*, ranging in length from 6 to 20 bases. These motifs are composed of 65,747 unique, non-coding, conserved exemplar sites covering 566,666bp (~0.8% of the *C. elegans* input sequence). These sites are distributed across all non-coding regions but have the most occurrences in the promoter regions, as would be expected for regulatory sites (Table 3). We make these motifs available as position-specific count matrices at http://ural.wustl.edu/~nihuegbu/Magma/homepage.html.

Table 3: Distribution of exemplar sites in different non-coding sequence classes

Location	Number of Sites	Coverage (bp)	Size of input region (bp)	Fraction of input region
2kb 5' Intergenic	34,278	258,322	21,532,733	1.20%
5'UTR	2,596	15,411	461,624	3.34%
1st Intron	15,514	73,904	7,918,585	0.93%
Other Intron	27,787	122,333	23,691,626	0.52%
3'UTR	4,436	27,111	1,934,557	1.40%

* Note: Some of these sites overlap different regulatory regions of multiple genes

Evaluation of Magma *C. elegans* motifs

We assessed whether Magma's motifs are consistent with the known binding sites for the few characterized factors and with other information about regulatory interactions. Because we do not expect Magma's exemplar sites for each motif to be a comprehensive list of all sites for its associated TF, we scan each non-coding region in our input with the PWM for each motif to determine if it was significantly enriched in instances of the motif. The expected number of motif instances arising by chance is determined by the information content of the motif (Schneider, Stormo et al. 1986; Hertz and Stormo 1999), while the observed number is the actual number of sites within each dataset whose score exceeds the information content of the motif. The score of a putative motif with respect to a given dataset is the log-likelihood ratio

$$LLR(motif \mid dataset) = observed \ln \frac{observed}{expected}.$$

One of the best-characterized TFs in *C. elegans* is the Nuclear Factor I (NFI). Whittle, Lazakovitch et al. (2009) performed ChIP-CHIP for NFI, probing its *in vivo* targets at 55 regions (~1500bp each). Magma finds two motifs that are strongly enriched within those regions, both very similar to the known consensus of $TTGGCAN_3TGCCAA$ (Figure 11).

Figure 11: Log Likelihood Ratio uncovers NFI-like motifs on NFI ChIP peaks.
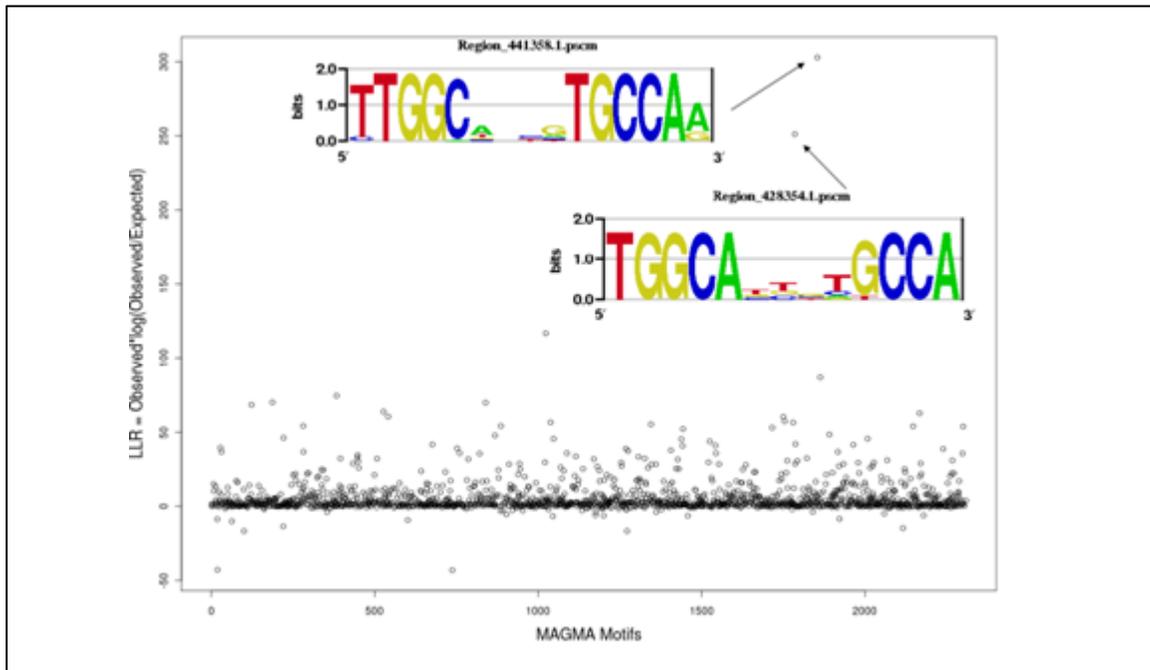
The modENCODE consortium identified regions from ChIP-Seq experiments that bound several TFs (Gerstein, Lu et al. 2010). These regions, with average length of 200 bases, were filtered to remove those that overlapped ubiquitous HOT sites, leaving 74,065 regions from 28 samples that bound a total of 23 different TFs (PHA-4 was assayed at 6 different developmental and environmental conditions). For each sample, we ranked the motifs using the above LLR score. For the three TFs with known motifs, the most significant Magma motif matches the known consensus (Table 4; for the PHA-4-YA set the second-ranked motif matches the consensus). Significant motifs were found for each of the remaining ChIP-Seq datasets, but since the TFs binding these sites have unknown motifs, we could not use them to validate Magma's performance.

Table 4: Magma motifs in modENCODE ChIP peaks

ChIP-Seq Sample	Class	TF	Known Specificity	Magma Motif LOGO	LLR Rank
HLH1_EMB	bHLH	HLH-1	E-Box (CANNTG)	Region_51609.1.pscm	1
PHA4_EMB	Forkhead	PHA-4	TRTTKRY	Region_236436.1.pscm	1
PHA4_L1					1
PHA4_L2					1
PHA-4-Late_Emb					1
PHA-4-Starved_L1					1
PHA-4-YA					2
ELT3_L1	GATA-Zn Finger	ELT-3	GATA-site	Region_87102.1.pscm	1

We also identified significant motifs for 12 factors with at least 10 promoter binding observations from the EDGE database of Yeast-One Hybrid (Y1H) experiments (Barrasa, Vaglio et al. 2007), though again the correct motifs for these sites are not known *a priori*. The Oreganno database lists187 different experimentally tested binding sites and *cis*-regulatory modules in *C. elegans* (Montgomery, Griffith et al. 2006; Griffith,

Montgomery et al. 2008), which includes the annotated bound factors for several sites. We find significant matches among our Magma motifs for 185 of these sites, including motifs whose specificity resembles that of TFs matching annotated PHA-4, ELT-2, and DAF-19 sites.

Hunt-Newbury and colleagues built promoter:GFP fusion libraries for approximately 2000 *C. elegans* genes (Hunt-Newbury, Viveiros et al. 2007) and cataloged the temporal and spatial expression of the green fluorescent protein. Chikina and colleagues (Chikina, Huttenhower et al. 2009) used support vector machines (SVMs) to predict other genes from *C. elegans* with similar expression profiles and achieved 90% precision for all of the major tissues (intestine, hypodermis, muscle, neurons, pharynx) except germ-line. Using these two datasets (the gold GFP dataset and the SVM predictions), we identified enriched motifs by computing an occupancy score for each motif and each 1kb-promoter in each tissue-specific gene set (Granek and Clarke 2005). We recovered several known *cis*-regulatory elements that regulate or establish tissue expression. For instance, ELT-2 is a zinc finger protein that is known to bind to GATA *cis*-based elements to regulate transcription in *C. elegans* intestines (McGhee, Sleumer et al. 2007). Figure 12 shows three GATA motifs and their tissue enrichments (log p-values). Although GATA-elements are mostly enriched in the promoters of intestine-expressed genes, we also found it enriched in the introns (especially the first intron) of neuronal and muscle tissue-types such as pharynx, uterus, and vulva, consistent with previous developmental studies highlighting the broad role of GATA-factors in development (Spencer, Zeller et al. 2011). We re-discovered other known *cis*-acting elements that endow tissue-specific expression, such as PHA-4- and PHA-4-variant-like motifs enriched in the pharynx.

83

Figure 12: Magma GATA-like motifs are mostly enriched in intestinal cells
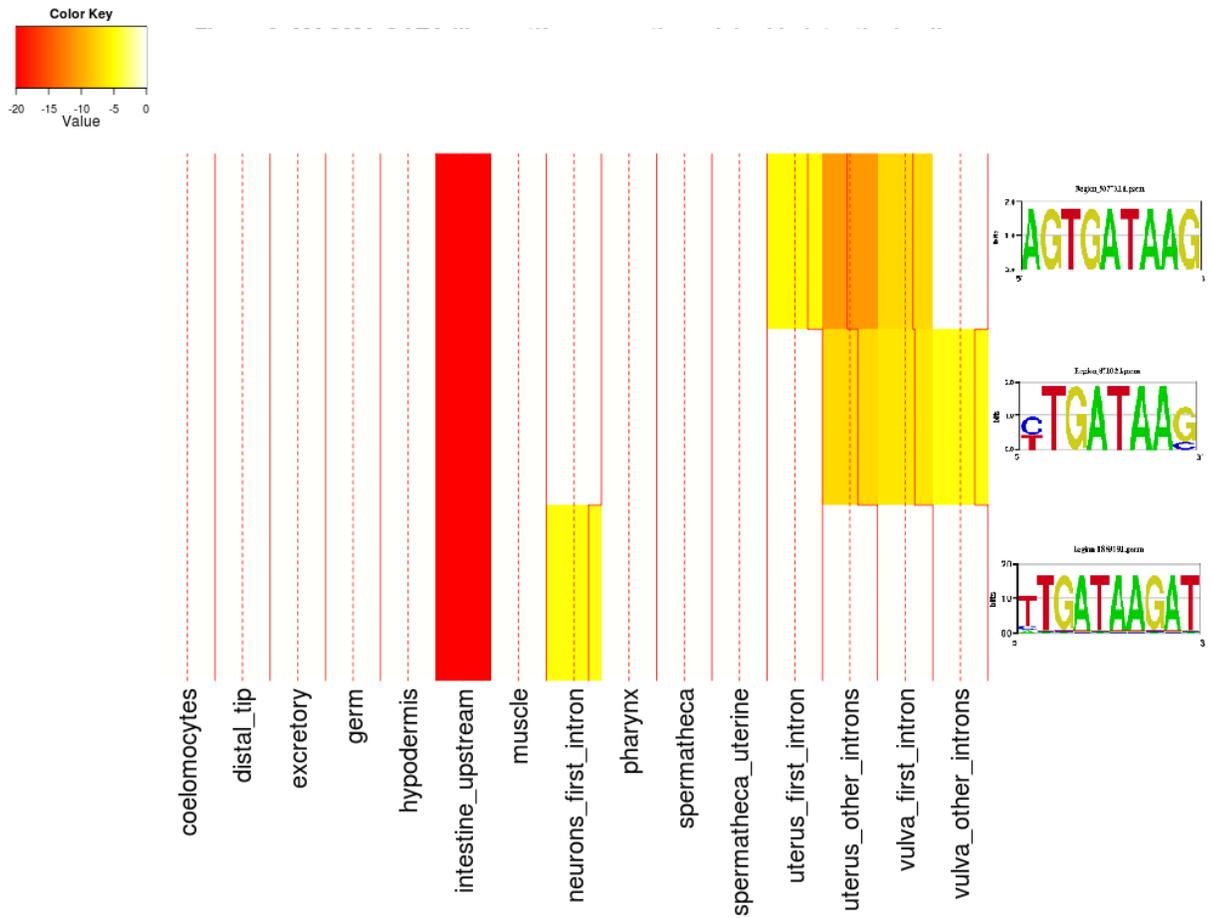
We further analyzed 88 *C. elegans* ChIP and expression microarray series data sets from the GEO Omnibus database, including 1,362 total samples. Similarly to the previous section, we analyzed the occupancy scores for our discovered motifs to uncover significant enrichments with the differentially regulated genes from each expression sample. We identified significant motifs for 991 different samples. We found that a motif matching the known specificity of Daf-16 (GTTGTTTAC) is significantly enriched in

daf-2/daf-16 mutant experiments (McElwee, Schuster et al. 2004). Daf-16 has also been shown to be involved in starvation response in *C. elegans* (Henderson and Johnson 2001), and samples from starvation experiments (Baugh, Demodena et al. 2009), are significantly enriched for the same motif.

Discussion and Conclusions

We have described Magma, a program that identifies motifs that are conserved across species and occur in several locations within the reference genome. In a comparison to the PhyloNet program on the yeast genome, we found slightly higher sensitivity with greatly increased speed, about 70x faster. The entire non-coding conserved genome of *C. elegans*, about 70Mbp, can be analyzed in less than four hours on a single CPU. We observed that Magma scales sub-quadratically with its input size, due to lower density of strongly conserved regions hence less HSP extensions per seed. Although the lack of extensive knowledge about regulatory motifs in *C. elegans* hinders a comprehensive evaluation of Magma's specificity, comparison to known motifs from a variety of experimental datasets show that its motifs are generally consistent with existing knowledge. Finally, we posit that these motifs likely represent specificities for TFs involved in various regulatory networks controlling gene expression in different conditions and developmental processes.

Chapter 4: Discovering *cis*-Regulatory Modules in

C. elegans using Magma motifs

Cis-regulatory modules are comprised of clustered transcription factor binding sites

A cis-regulatory module (CRM) is a segment of DNA that contains clustered transcription factor binding sites which function together to regulate the particular expression patterns of the associated gene. Many studies have shown that in higher organisms, CRM is a common strategy in regulating gene expression. Clustered TF binding sites direct tissue/temporal-specific *in vivo* gene expression (Kirchhamer, Yuh et al. 1996; Arnone and Davidson 1997; Wasserman and Sandelin 2004; Blanchette, Bataille et al. 2006; Sinha, Liang et al. 2006). Consequently, clusters of TF binding sites, along with phylogenetic conservation and other measures of "regulatory potential", have been widely used in computational prediction of CRMs and is a more reliable indicator of *in vivo* regulatory function of DNA sequences (Kolbe, Taylor et al. 2004; Wasserman and Sandelin 2004; King, Taylor et al. 2005; Blanchette, Bataille et al. 2006; Sinha, Liang et al. 2006; Taylor, Tyekucheva et al. 2006; Ferretti, Poitras et al. 2007) .

Due to the speedup modifications, Magma is capable of predicting regulatory sites from a larger space which encompasses more genes than the earlier discovered PhyloNet motifs in Chapter 2. Additionally, due to a better redundancy reduction algorithm, Magma motifs are more specific. They are less redundant, do not have overlapping exemplar sites, and are fewer than the PhyloNet results. Consequently, we inquired if these Magma sites cluster like known transcription factor binding sites. Conserved, exemplar instances from Magma-predicted motifs are highly clustered along the regulome. Furthermore, they occur more frequently than expected within known cis-regulatory modules (p<1e-9) (See Figure 13). Based on this, I extended the earlier framework (in Chapter 2) to predict

CRMs based on this more comprehensive motif collection.

Figure 13: Magma exemplar sites are clustered within known cis-regulatory modules

The "Known CRMs" track describes the genome intervals that have been validated by previous studies to have regulatory capacity; The tracks starting with the prefix "Region_" under the Magma title are conserved exemplar sites for the different Magma motifs discovered in this region; "RefSeq Genes" describe the gene models from the RefSeq database; and the "Conservation" and other genomes below it describe the conservation of this region in other nematode genomes.
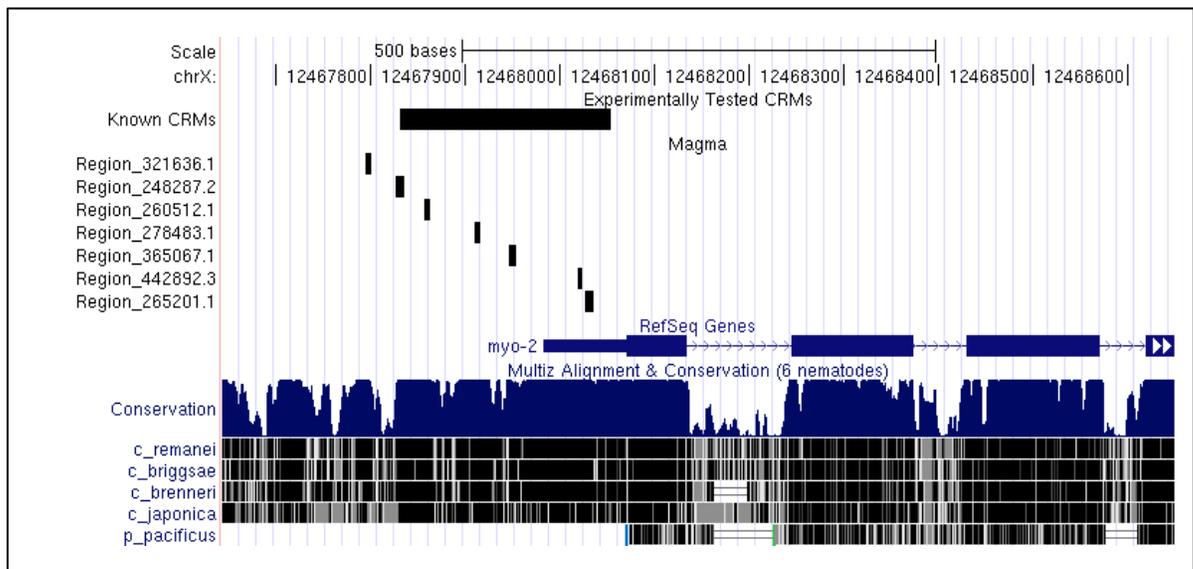
Predicting cis-regulatory modules

Predicting cis-regulatory modules from sequence alone is challenging because a

regulome-wide scan tends to over predict putative binding sites. Consequently this leads

to an exaggeration of predicted clustered sites or CRMs (specificity problem). On the

other hand, too few CRMs are predicted if only conserved exemplar sites are used as not

all functional elements are conserved with enough instances to meet the multi-gene multi-

species constraints (sensitivity problem). To ameliorate these seemingly-competing

methods, I designed a framework that combines the two approaches to exploit both the

sensitivity of regulome-scanning and the specificity of conserved exemplar sites.

Regulome-scanning methods over-predict putative sites and CRMs for several reasons.

These include: (1) use of excessive (possibly thousands) predicted motifs; (2) use of non-

complex and degenerate motifs to score a large space; (3) a significant portion of the

motifs are redundant; (4) overly relaxed site-matching criteria; (5) overly generous site

extension criteria to define modules. The method presented here reduces these unwanted

results by using the previously discovered collection of 2309 Magma motifs, a reasonably

sized set given the ~900 estimated *C. elegans* TFs and many other RNA-binding proteins

& RNAs. Additionally, these motifs have little redundancy due to our post-processing

set-cover step. We further ameliorate the problem of over-predicting CRMs by only

scanning motifs with 12 or more bits of information using Patser (Hertz and Stormo

1999) and find peaks using CERMOD (Zhao, Ihuegbu et al. 2011). CERMOD calculates

the average number of Patser-predicted binding and a Z-score for each position along a

sequence. Using Z-scores $\geq$ 3.09 (corresponding to p-value = 1.0E-3), CERMOD selects

peaks and extends them in both directions if the next position with Z-score > 0 is within

30bp. We keep scanned peaks that are comprised of 3 or more different motif instances

and combine the intervals with the original high-confident conserved Magma exemplar

sites from all 2309 motifs. After clustering nearby sites and peaks (within 75bp), we

define our final set of predicted CRMs. We predict 110,933 Magma-based CRMs

covering ~9.75Mbp of intergenic and intronic bases (an average of ~88bp). Table 5

shows the distribution of predicted CRMs across the regulome near coding genes.

Consistent with known regulatory elements, the upstream and first intron regions contain the greatest density of predicted modules.

Table 5: Distribution of Predicted CRMs surrounding protein-coding genes

Location	Number of CRMs	Coverage (bp)	Size of input region (bp)	Fraction of input region
1kb 5' Intergenic	51,567	4,575,173	21,043,726	21.74%
5' UTR	3,243	394,890	9,080,734	4.35%
1st Intron	20,149	1,922,953	11,952,010	16.09%
Other Intron	45,576	4,014,064	25,204,976	15.93%
3'UTR	5,124	403,352	3,920,828	10.29%

* Note: Some of these sites overlap different regulatory regions of multiple genes

These predicted CRMs are available at the site http://ural.wustl.edu/~nihuegbu/Magma/homepage.html as tracks that can be viewed in the UCSC Browser. This presentation forum is especially useful for viewing the predicted CRMs in context with the original Magma conserved sites, gene models, conservation, and other relevant information.

Evaluating predicted cis-regulatory modules (CRMs)

To evaluate our predicted CRMs, we performed a thorough search of existing literature and the Oreganno database (Montgomery, Griffith et al. 2006; Griffith, Montgomery et al. 2008) of regulatory elements to identify *C. elegans* genes which have nearby non-coding regions that have been analyzed for regulatory capacity. We identified 44 genes that contain 79 CRMs of 400bp-length or less (see Table 6 at the end of this chapter). Sixty-nine (69) of these CRMs are in upstream regions and 10 are in intronic regions. These experimentally determined regulatory regions are important for their corresponding gene expression at various developmental stages and in a broad range of tissues including: neurons, hypoderms, excretory cells, muscle precursor cells, adult muscle cells, vulva cells, sheath cells, etc. They were determined by deletion and/or enhancer assays. Wherever possible, we use regions that are determined by enhancer assay because it better defines the boundary of regulatory regions that are sufficient in regulation.

The predicted CRMs overlap 51 of the 79 (~65%) experimentally defined modules. Since most of the known CRMs reside in the upstream regions of genes (69/79) we evaluated the performance of our upstream predicted CRMs against this set. Forty-four (44) of the 69 known upstream CRMs overlap a predicted upstream CRM. It is difficult to assess the significance of this overlap since the predicted CRMs and the known CRMs often do not have comparable lengths. Larger portions of non-coding sequences are usually tested in enhancer or deletion assays to curb the risk of failure. Hence the reported CRM intervals are often much larger than the actual functioning portions. Additionally, enhancer assays are laborious (a major reason why only few CRMs are known). Therefore many non-

coding portions which are predicted to be CRMs may actually regulate nearby genes but have not been tested. Nevertheless, we see that this overlap in upstream predicted CRMs and known CRMs is significant at $p \ll 0.001$ using a Chi-Square test (see Figure 14).

Figure 14: Predicted Upstream CRMs significantly overlap known Upstream CRMs

Entire Promoterome
(~20.70Mbp)

Predicted Upstream CRMs
(~4.84Mbp)

Known Upstream CRMs
(14,499bp)

~3.9kbp
$p \ll 0.001$

Experimental test of CRM prediction

mlc-1 and *mlc-2* are the two muscle regulatory myosin light chain genes in *C. elegans*.

They are divergently located and share a 2.6 kb intergenic region. It was shown that they

are both expressed in the body-wall muscles, pharyngeal muscles, and vulval muscles.

However, the intergenic region has not been analyzed in detail to identify all the

regulatory sequences that drive their expression. Previous study has shown that the first

400 bp of *mlc-2* upstream sequence is enough to drive its expression in the body wall

muscle cells (GuhaThakurta, Schriefer et al. 2004). To gain better information about

transcriptional regulation of *mlc-1* and *mlc-2*, the new Magma-based module prediction

method was applied on their intergenic region and predicted modules were

experimentally tested. Three (3) Magma-based CRMs were predicted within this 2662 bp

DNA fragment (Figure 15): [17,456,614 - 17,456,944] (nearest to *mlc-2*), [17,457,025 -

17,457,240] (near *mlc-2*) and [17,458,974 - 17,459,154] (nearest to *mlc-1*). The predicted

CRM closest to *mlc-2* overlaps a 400bp region that had been previously shown to drive

expression in the body wall muscle (GuhaThakurta, Schriefer et al. 2004). As tested in

Chapter 2, the DNA fragment which covers the middle predicted module showed

enhancer activity for pharyngeal muscle. Again, it is interesting to note that *mlc-2* gene is

known to be expressed in both body wall and pharyngeal muscle. We have uncovered a

separated CRM that specifically drives its pharyngeal muscle expression.

Figure 15: Verifying a Magma-based module for pharyngeal muscle

The "Magma" track contains conserved exemplar sites for the different Magma motifs discovered in this region; The track below the title "Experimentally Tested Regions" are the two regions tested in Chapter 2; The "Known Modules" track describes the genome intervals that have been validated by previous studies to have regulatory capacity; The "Magma CRMs" track shows the predicted CRMs using Magma motifs in this chapter; The track below "RefSeq Genes" shows the gene models from the RefSeq database; and the "Conservation" and other genomes below it describe the conservation of this region in other nematode genomes.
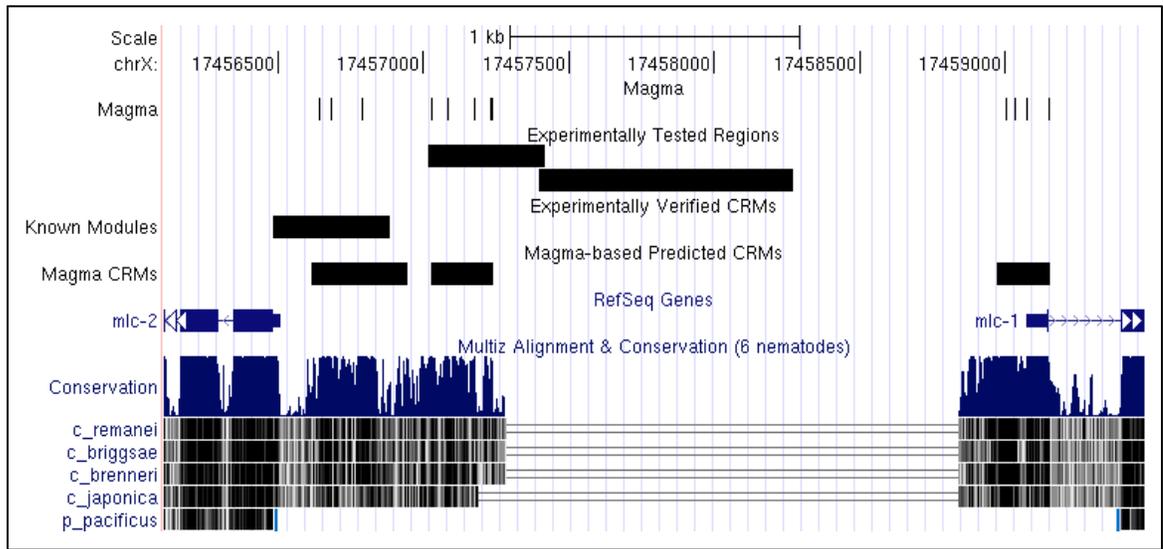
Conclusion

Cis-Regulatory modules (CRMs) are DNA stretches of dense clustered transcription factor binding sites that independently endow a nearby gene with conditions-specific spatial/temporal expression patterns. Likewise, the previously discovered Magma exemplar sites were observed to be highly clustered, especially within known modules. This observation spurred the development of a framework for predicting CRMs in *C. elegans* to augment the few known modules.

This approach ameliorates the challenges of CRM prediction: high sensitivity but over-prediction with scanning-only methods and low sensitivity but high specificity with conservation-only methods. The proposed solution combines both scanning and conservation information. The regulome is scanned with more complex motifs (thereby reducing over-prediction due to degenerate or simple motifs) and their instances are clustered with nearby conserved exemplar sites into putative CRM windows.

This method predicts 110,933 Magma-based CRMs covering ~9.75Mbp of intergenic and intronic bases (an average of ~88bp). These predictions are evaluated by measuring their overlap with known, literature compiled CRMs. Predicted CRMs overlap with 51/79 (~65%) of known CRMs. Furthermore upstream predicted CRMs overlap 44 of the known 69 upstream CRMs significantly at $p \ll 0.0001$.

Finally, two regions upstream to *mlc-2* are experimentally tested: one includes a predicted CRM and the other does not. We show that the region that includes the predicted CRM drives pharyngeal expression of *mlc-2* and no enhancer activity was observed for the other.

Table 6: Experimentally tested cis-regulatory modules

Chr	Start	End	Name	Expression	Reference
chrI	3682153	3682215	ftn-2	ion dependent transcription in intestine (necessary and sufficient)	An iron enhancer element in the FTN-1 gene directs iron-dependent expression in Caenorhabditis elegans intestine. Romney SJ, Thacker C, Leibold EA. J Biol Chem. 2008;283(2):716-25 , necessary and sufficient
chrI	7267057	7267434	myo-1		
chrI	14142193	14142462	kal-1	AIY, other neurons (EA).	Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in C. elegans.
chrI	14863499	14863679	unc-54		Sequence requirements for myosin gene expression and regulation in Caenorhabditis elegans. Okkema PG, Harrison SW, Plunger V, Aryana A, Fire A. Genetics. 1993;135(2):385-404.
chrII	9614736	9614976	hlh-6	pharyngeal glands, minimal promoter	Gland-specific expression of C. elegans hlh-6 requires the combinatorial action of three distinct promoter elements. Raharjo I, Gaudet J. Dev Biol. 2007;302(1):295-308.

chrII	10316142	10316447	gcy-5	neurons (EA or minimal promoter)	The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron. Etchberger JF, Lorch A, Sleumer MC, Zapf R, Jones SJ, Marra MA, Holt RA, Moerman DG, Hobert O. Genes Dev. 2007 Jul 1;21(13):1653-74.
chrII	10912854	10912950	sra-11	AIY, other neurons(EA).	
chrIII	5934897	5935197	zmp-1	AC, VulA, VulE (EA)	cis-Regulatory control of three cell fate-specific genes in vulval organogenesis of Caenorhabditis elegans and C. briggsae. Kirouac M, Sternberg PW. Dev Biol. 2003;257(1):85-103
chrIII	7541988	7542326	lin-39	p7/8, p5/6, VCN, p5.p, p8.p (EA)	
chrIII	7757896	7758128	cdh-3	AC (EA)	
chrIII	10186156	10186394	unc-47	minimal promoter	Coordinated transcriptional regulation of the unc-25 glutamic acid decarboxylase and the unc-47 GABA vesicular transporter by the Caenorhabditis elegans UNC-30 homeodomain protein. Eastman C, Horvitz HR, Jin Y. J Neurosci. 1999;19(15):6225-34. minimal promoter

chrIV	5537956	5538014	fem-1	minimal promoter	Gaudet J, VanderElst I, Spence AM. Post-transcriptional regulation of sex determination in Caenorhabditis elegans: widespread expression of the sex-determining gene fem-1 in both sexes. Mol Biol Cell. 1996 Jul;7(7):1107-21.
chrV	7548176	7548238	ftn-1	ion dependent transcription in intestine (necessary and sufficient)	An iron enhancer element in the FTN-1 gene directs iron-dependent expression in Caenorhabditis elegans intestine. Romney SJ, Thacker C, Leibold EA. J Biol Chem. 2008;283(2):716-25 , necessary and sufficient
chrV	10271236	10271379	snap-25	motor neurons (necessary and sufficient)	Neuron cell type-specific SNAP-25 expression driven by multiple regulatory elements in the nematode Caenorhabditis elegans.
chrV	10270425	10270561	snap-25	amphid, phasmid neurons (necessary and sufficient)	
chrV	10921734	10922045	nas-31	exlusive excretory cell (minimal promoter)	Distinct regulatory elements mediate similar expression patterns in the excretory cell of Caenorhabditis elegans. Zhao Z, Fang L, Chen N, Johnsen RC, Stein L, Baillie DL. J Biol Chem. 2005, 18;280(46):38787-94.

chrX	7118491	7118844	hen-1	AIY, other neurons (EA).	Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in C. elegans.
chrX	7175347	7175668	bar-1	subset of ventral cord neurons (EA)	
chrX	7537858	7538018	dpy-7	hypodermal cell (EA)	cis regulatory requirements for hypodermal cell-specific expression of the Caenorhabditis elegans cuticle collagen gene dpy-7.
chrX	15298146	15298427	ser-2	minimal promoter	Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in C. elegans.
chrX	17456484	17456883	mlc-2	muscle (minimal promoter)	Novel transcription regulatory elements in Caenorhabditis elegans muscle genes. GuhaThakurta D, Schriefer LA, Waterston RH, Stormo GD. Genome Res. 2004 Dec;14(12):2457-68.
chrI	6780601	6780819	cye-1	neurons, intestin	Brodigan, T.M., Liu, J., Park, M., Kipreos, E.T., and Krause, M. (2003). Cyclin E expression during development in Caenorhabditis elegans. Dev. Biol. 254, 102.115.
chrI	6780211	6780297	cye-1	seam cell, embryonic	
chrI	6780085	6780211	cye-1	vulval and p	
chrII	4518519	4518598	hlh-1	BWM (EA)	use enhancer assay position instead of deletion assay position

chrII	4518106	4518330	hlh-1	D, MS (DA)	Elements regulating cell- and stage-specific expression of the C. elegans MyoD family homolog hlh-1. Krause M, Harrison SW, Xu SQ, Chen L, Fire A. Dev Biol. 1994;166(1):133-48.
chrII	4517123	4517476	hlh-1	mature BWM (DA)	
chrII	4516585	4516939	hlh-1	C, MS (DA)	
chrII	4516450	4516518	hlh-1	MS-grandaughter (DA)	
chrII	4516245	4516422	hlh-1	BWM (EA)	
chrII	10217957	10218110	ref-1	endodermal cell (EA)	Notch-GATA synergy promotes endoderm-specific expression of ref-1 in C. elegans. Neves A, English K, Priess JR. Development. 2007 Dec;134(24):4459-68. Epub 2007 Nov 14.
chrIII	7552979	7553249	ceh-13	mail tail (necessary and sufficient)	Expression of the C. elegans labial orthologue ceh-13 during male tail morphogenesis. Stoyanov CN, Fleischmann M, Suzuki Y, Tapparel N, Gautron F, Streit A, Wood WB, MÃ¼F. Dev Biol. 2003 Jul 1;259(1):137-49.
chrIII	7810911	7811214	egl-5	v6 lineage (enhancer assay)	Dissection of cis-regulatory elements in the C. elegans Hox gene egl-5 promoter. Teng Y, Girard L, Ferreira HB, Sternberg PW, Emmons SW. Dev Biol. 2004 ,15;276(2):476-92.
chrIII	7808959	7809273	egl-5	tail hyperderm, sex muscle (enhancer assay)	

chrIII	7838410	7838728	ceh-23	CAN (EA)	
chrIII	7837066	7837250	ceh-23	neurons (EA)	
chrIII	7832700	7832743	ceh-23	neurons (EA)	
chrIII	12942359	12942538	unc-25	minimal promoter	Coordinated transcriptional regulation of the unc-25 glutamic acid decarboxylase and the unc-47 GABA vesicular transporter by the Caenorhabditis elegans UNC-30 homeodomain protein. Eastman C, Horvitz HR, Jin Y. J Neurosci. 1999;19(15):6225-34.
chrIV	11057629	11057783	lin-3	anchor cell (EA)	A cell-specific enhancer that specifies lin-3 expression in the C. elegans anchor cell for vulval development. Hwang BJ, Sternberg PW. Development. 2004;131(1):143-51.
chrV	8387145	8387332	gcy-7	neurons (EA or minimal promoter)	The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron. Etchberger JF, Lorch A, Sleumer MC, Zapf R, Jones SJ, Marra MA, Holt RA, Moerman DG, Hobert O. Genes Dev. 2007 Jul 1;21(13):1653-74.
chrV	10671809	10671955	ceh-22	enhance expression (DA)	

chrV	13811391	13811786	ceh-24	m8 (DA)	Muscle and nerve-specific regulation of a novel NK-2 class homeodomain factor in Caenorhabditis elegans. Harfe BD, Fire A. Development. 1998 Feb;125(3):421-9.
chrV	13811376	13811493	ceh-24	m8 (DA)	
chrV	13810684	13810741	ceh-24	head neurons (DA)	
chrX	489472	489793	egl-17	early expression (EA)	Cis regulatory requirements for vulval cell-specific expression of the Caenorhabditis elegans fibroblast growth factor gene egl-17. Cui M, Han M. Dev Biol. 2003; 257(1):104-16.
chrX	489428	489491	egl-17	VulD, VulC (EA)	vulC, vulD
chrX	488222	488385	egl-17	early stage, high level expression (EA)	
chrX	487205	487561	egl-17	M4 cell (DA)	deletion assay
chrX	489651	489753	egl-17	vulE, vulF,	cis-Regulatory control of three cell fate-specific genes in vulval organogenesis of Caenorhabditis elegans and C. briggsae. Kirouac M, Sternberg PW. Dev Biol. 2003;257(1):85-103.
chrX	489402	489560	egl-17	vulC, vulD	
chrX	1074724	1074931	lim-6	neurons (EA or minimal promoter)	The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron. Etchberger JF, Lorch A, Sleumer MC, Zapf R, Jones SJ, Marra MA, Holt RA, Moerman DG, Hobert O. Genes Dev. 2007 Jul 1;21(13):1653-74.

chrX	5100763	5101009	vit-2		MacMorris, M., Broverman, S., Greenspoon, S., Lea, K., Madej, C., Blumenthal, T., and Spieth, J. (1992). Regulation of vitellogenin gene expression in transgenic Caenorhabditis elegans: short sequences required for activation of the vit-2 promoter. Mol. Cell. Biol. 12, 1652.1662
chrX	8116264	8116578	hlh-8	minimal promoter	Analysis of a Caenorhabditis elegans Twist homolog identifies conserved and divergent aspects of mesodermal patterning. Harfe BD, Vaz Gomes A, Kenyon C, Liu J, Krause M, Fire A. Genes Dev. 1998 Aug 15;12(16):2623-35.
chrX	12467384	12467700	myo-2		Sequence requirements for myosin gene expression and regulation in Caenorhabditis elegans. Okkema PG, Harrison SW, Plunger V, Aryana A, Fire A. Genetics. 1993;135(2):385-404.
chrX	12467306	12467612	myo-2		
chrX	12467831	12468053	myo-2	minimal promoter	
chrX	14684459	14684534	csq-1	minimal promoter	Analysis of calsequestrin gene expression using green fluorescent protein in Caenorhabditis elegans. Cho JH, Eom SH, Ahnn J. Mol Cells. 1999;9(2):230-4.
chrX	14684269	14684480	csq-1	BWM (EA)	

chrX	16367009	16367296	ace-1	minimal promoter	Structure and promoter activity of the 5' flanking region of ace-1, the gene encoding acetylcholinesterase of class A in Caenorhabditis elegans. Culetto E, Combes D, Fedon Y, Roig A, Toutant JP, Arpagaus M. J Mol Biol. 1999;290(5):951-66.
chrX	16366599	16366804	ace-1	pm5, neuron (DA)	
chrX	16365248	16365566	ace-1	BWM, anal muscle, vulval muscle cells (DA)	
chrX	16364899	16365248	ace-1	BWM, anal muscle (DA)	
chrI	14862280	14862368	unc-54		Sequence requirements for myosin gene expression and regulation in Caenorhabditis elegans. Okkema PG, Harrison SW, Plunger V, Aryana A, Fire A. Genetics. 1993;135(2):385-404.
chrV	10265894	10266133	snap-25	motor neurons (EA)	Neuron cell type-specific SNAP-25 expression driven by multiple regulatory elements in the nematode Caenorhabditis elegans.
chrV	10263906	10263934	snap-25	motor neurons (EA)	
chrV	10265323	10265348	snap-25	pharyngeal neurons (EA)	
chrV	10263829	10263858	snap-25	mechanosensoty neruons (EA)	

chrII	4519082	4519125	hlh-1	MS-granddaughter cells	Elements regulating cell- and stage-specific expression of the C. elegans MyoD family homolog hlh-1. Krause M, Harrison SW, Xu SQ, Chen L, Fire A. Dev Biol. 1994;166(1):133-48.
chrII	4519751	4519862	hlh-1	GLR cells (EA)	
chrV	10671979	10672225	ceh-22		OREG0001740*
chrV	13502235	13502489	avr-15		OREG0001745*
chrX	2215627	2215881	peb-1		OREG0001746*
chrX	15517456	15517710	eat-20		OREG0001747*
chrV	6691816	6692212	mtl-1		OREG0001824*
chrV	14018598	14018950	mtl-2		OREG0001825*
chrX	7175658	7175978	bar-1		OREG0001987*
chrX	489420	489562	egl-17		OREG0002003*
chrX	489654	489755	egl-17		OREG0002011*
chrIII	7754894	7755048	cdh-3		OREG0002021*

* Identifier in Oreganno Database of regulatory elements

Chapter 5: HLH-30 is a novel transcription factors involved in host defense response[3]

Introduction

As well as being an excellent tool for genetic manipulation, *C. elegans* has also been extensively used as a model system to study many human pathogens and infections (Kurz and Ewbank 2000; Aballay and Ausubel 2002; Couillault and Ewbank 2002; Sifri, Begun et al. 2005; Powell and Ausubel 2008). Many of these cause potent intestinal infections in nematodes that result in death. Like in humans, the major site for host-microbe interaction is along the intestinal tract. Yet unlike humans, *C.elegans* intestinal tract is made up of 20 intestinal epithelial cells (IECs) that are non-reneweable; they do not shed and proliferate like mammalian IECs (Figure 16).

Figure 16: Schematic representation of the C. elegans intestine

The following figure and legend were taken from (Irazoqui, Urbach et al. 2010). The *C. elegans* intestine is composed of 20 intestinal epithelial cells. These cells are organized in 9 rings: ring 1 contains four cells and rings 2–9 contain two cells each. The apical surface of each of the intestinal epithelial cells forms the microvillar brush border and faces the intestinal lumen. The intestinal epithelium is the major interface of interaction between *C. elegans* and ingested microbes
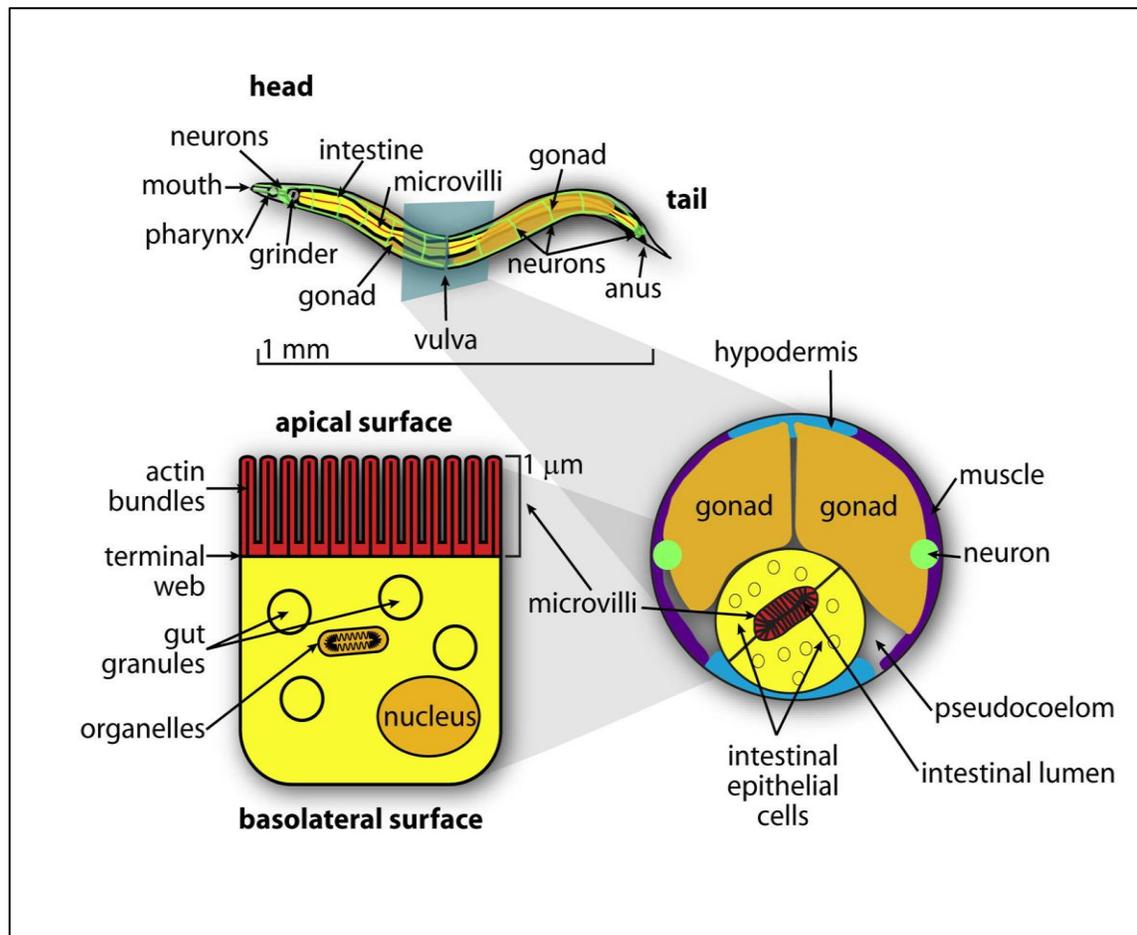
There are special advantages to using a transparent organism to study pathogenesis.

Animals can be infected, through their diet, with pathogens laced with reporter

constructs. These constructs contain a reporter that fluoresces as the poisoned diet

progresses along the pharynx and accumulates in the intestinal tracts (Irazoqui, Ng et al.

2008). Additionally pathogenesis can be monitored by plotting survival curves (Tan,

Mahajan-Miklos et al. 1999), observing morphological or behavioral changes (Hodgkin, Kuwabara et al. 2000; Pujol, Link et al. 2001; Zhang, Lu et al. 2005). Their gene expression differences can be ascertained by using microarrays, RNA-Seq and quantitative reverse polymerase chain reactions (qRT-PCR) (Troemel, Chu et al. 2006).

Staphylococcus aureus is a gram-positive bacterium that causes many diseases in animals (Sifri, Begun et al. 2003; Cuny, Friedrich et al. 2010). Furthermore increasing virulent methicillin-restistant strains are worrisome and motivating further studies into the host-immunity response to this pathogen (Boucher and Corey 2008). As reviewed in Irazoqui, Urbach et al. (2010), human colonization by *S. aureus* is widespread as 30% of the population carries the bacteria in the microflora of epithelia in the nasopharynx, skin, and intestine (Graham, Lin et al. 2006). *S. aureus* can also cause severe skin infections, osteomyelitis, endocarditis, food poisoning, pneumonia, and flesh-eating disease (Gordon and Lowy 2008). To successfully present these traits in the host, it deploys several virulence factors, including cytolysis which destroy the host's immune cells and tissues (Nizet 2007; Diep and Otto 2008).

Mechanisms of defense evolved before the split between invertebrates and vertebrates, thus many host signaling pathways are conserved and shared between nematodes and humans. Because nematodes represent a much simpler system, invertebrate genetic models have been used to identify conserved signaling pathways that also play key roles in mammalian innate immune response. Some of these include: p38 MAPK, insulin, TGF-β, and β-catenin pathways (Kurz and Ewbank 2003; Irazoqui, Ng et al. 2008; Zugasti and Ewbank 2009; Irazoqui, Urbach et al. 2010).

The pathogen *Staphylococcus aureus* causes intestinal pathogenesis and nematode

death, and has been shown to trigger a transcriptional host-defense response. Interestingly, this response is independent of the p38 MAPK, insulin and TGF-beta pathways, and relies only partially on β-catenin pathway (Irazoqui, Troemel et al. 2010). To further understand the host response to *S.aureus*, we embarked on this project to uncover the differential transcriptional regulation triggered in *C. elegans* in response to infection. Knowing the collection of genes differentially regulated by *S. aureus* may help uncover specific pathways activated by pathogenic infections in *C. elegans* and possibly in humans as well.

Results

An M-Box motif corresponding to HLH-30 is enriched in promoters of *S. aureus* induced genes

Previous published work measured differential expression due to *S. aureus* infection in worms with expression microarrays (Irazoqui, Troemel et al. 2010). Promoter regions for 688 genes were deemed significantly up-regulated after infection (p<0.05) and we searched for potential regulatory elements in them from a catalog of conserved motifs identified by Magma (Multiple Aligner of Genomic Multiple Alignments) (Ihuegbu, Stormo et al. 2011). Using *C. elegans* as the reference genome, Magma compares segments of the genome that are conserved across 5 other nematode species to identify conserved motifs that have many instances within the reference genome – typical characteristics of regulatory motifs. The catalog contains 2309 conserved motifs obtained from all classes of non-coding sequences: intergenic, intronic and UTR regions. Here we used the 2kb upstream region of the 688 up-regulated genes to identify which conserved motifs are significantly enriched (adjusted p<0.05, see Methods). After excluding promiscuous motifs that showed enrichment in many other unrelated conditions, two significant motifs were identified: a GATA element and an M-Box motif (Region_366771.1 see Figure 17). GATA elements have been previously associated with intestinal development and biology (Maduro 2006; Pauli, Liu et al. 2006; McGhee, Fukushige et al. 2009). Consequently, we decided to focus on the novel enrichment of the M-Box motif in this S. aureus up-regulated set (BH-corrected p<0.0056).

Figure 17: An M-Box motif is enriched in S. aureus induced genes
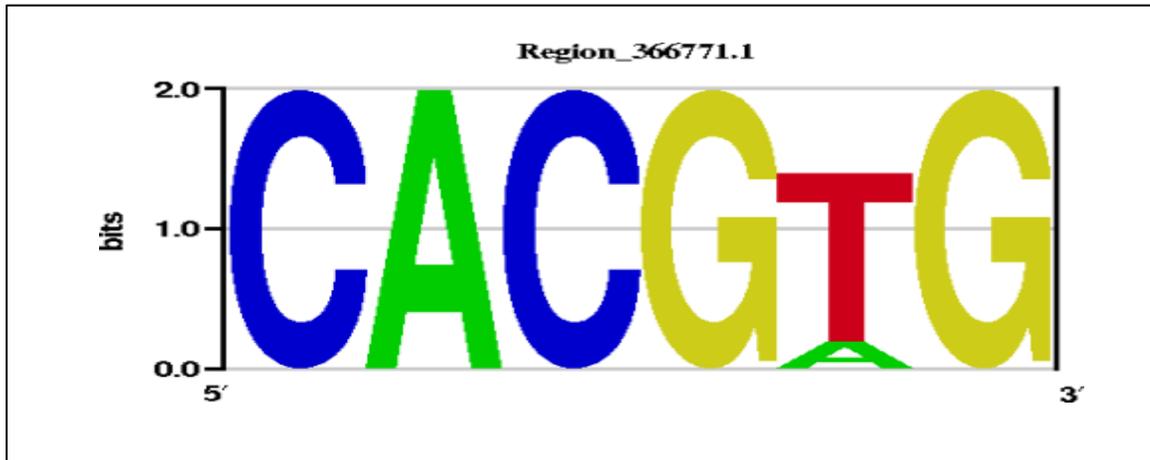
Grove and colleagues recently used Protein Binding Microarrays (PBM) to uncover the *in vitro* binding specificity of several bHLH proteins. They found three bHLH factors (including HLH-30) that specifically bind as a homodimer to CACGTG (Grove, De Masi et al. 2009). This is the exact consensus sequence of our discovered, enriched motif.

We searched the TRANSFAC database of previously discovered transcription factor binding site (TFBS) models from various organisms to find the closes match to our motif. We found that the discovered M-Box motif (CACGTG) is similar to a few mammalian HLH factors, including the MiTF/TFEB motif (CAT/CGTG). According to the KEGG database of orthologs and TreeFam families of proteins, Microphthalmia-associated transcription factor (MiTF) is a homolog of HLH-30 (Figure 18).

Figure 18: MiTF is homologous to HLH-30

The following figure is adapted from KEGG and TreeFam protein family trees
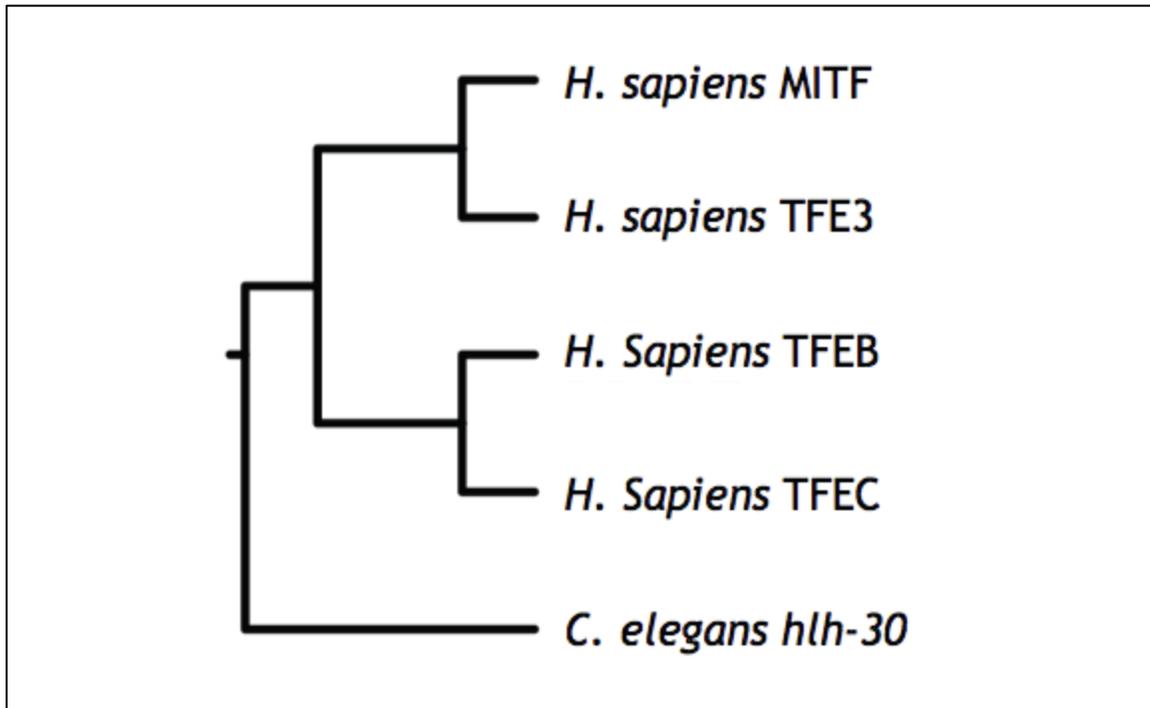
Out of the 3 bHLH proteins Grove, De Masi et al. (2009) showed to bind to the discovered M-Box consensus sequence, HLH-30 is the only protein that has a mammalian (specifically human) homolog (MiTF). Additionally, previous studies have implicated MiTF in a variety of stress responses (Saha, Singh et al. 2006; Liu, Fu et al. 2009).

All this led us to the hypothesis that the discovered M-Box motif is in fact a preferred binding profile for HLH-30 and that HLH-30 plays a significant role in response to *S. aureus* infection in nematodes. In accordance with this hypothesis, RT-qPCR experiments showed that *hlh-30* mRNA is up-regulated by 2 fold after 8 hours of infection (Figure 19).

Figure 19: *hlh-30* is up-regulated by 2-fold after 8 hrs of *S. aureus* infection

This figure displays the mRNA levels of *hlh-30* and *clec-71* (positive control) using qRT-PCR. At each time point, replicates for each strain (fed either *S. aureus* or a normal diet of *E. coli*), were collected and their mRNA levels for these genes were measured.
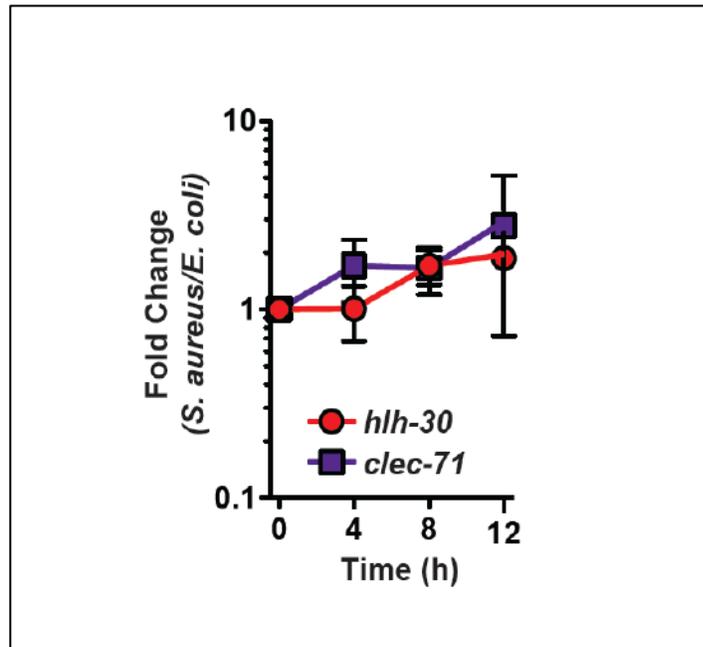
HLH-30 is localized in the nucleus upon S. aureus infection

To assess the localization of HLH-30 protein *in vivo*, we generated transgenic wild type animals expressing GFP- tagged HLH-30 under its own promoter (*hlh-30p*::HLH-30::GFP). We found that HLH-30 is expressed in many tissues, including the intestine – the major tissue exposed to the pathogen. At a sub-cellular level, HLH-30 equally localizes in the cytoplasm and the nucleus when animals are grown on standard non-pathogenic *E. coli* OP50. Thirty (30) minutes after transfer to *S. aureus* lawn, *hlh-30p*::HLH-30::GFP accumulates in the nucleus of 95% of exposed animals. This is illustrated by the punctate pattern in the "infected" panel of figure 20 versus the more

115

evenly distributed pattern in the "uninfected" panel. This is shows that HLH-30 localization is regulated during infection, supporting the hypothesis of its role as a major transcription factor involved in the host response. Altogether, these results demonstrate that HLH-30 is induced upon infection and targeted to the nucleus to trigger the host defense response.

Figure 20: HLH-30 is localized in the nucleus upon *S. aureus* infection
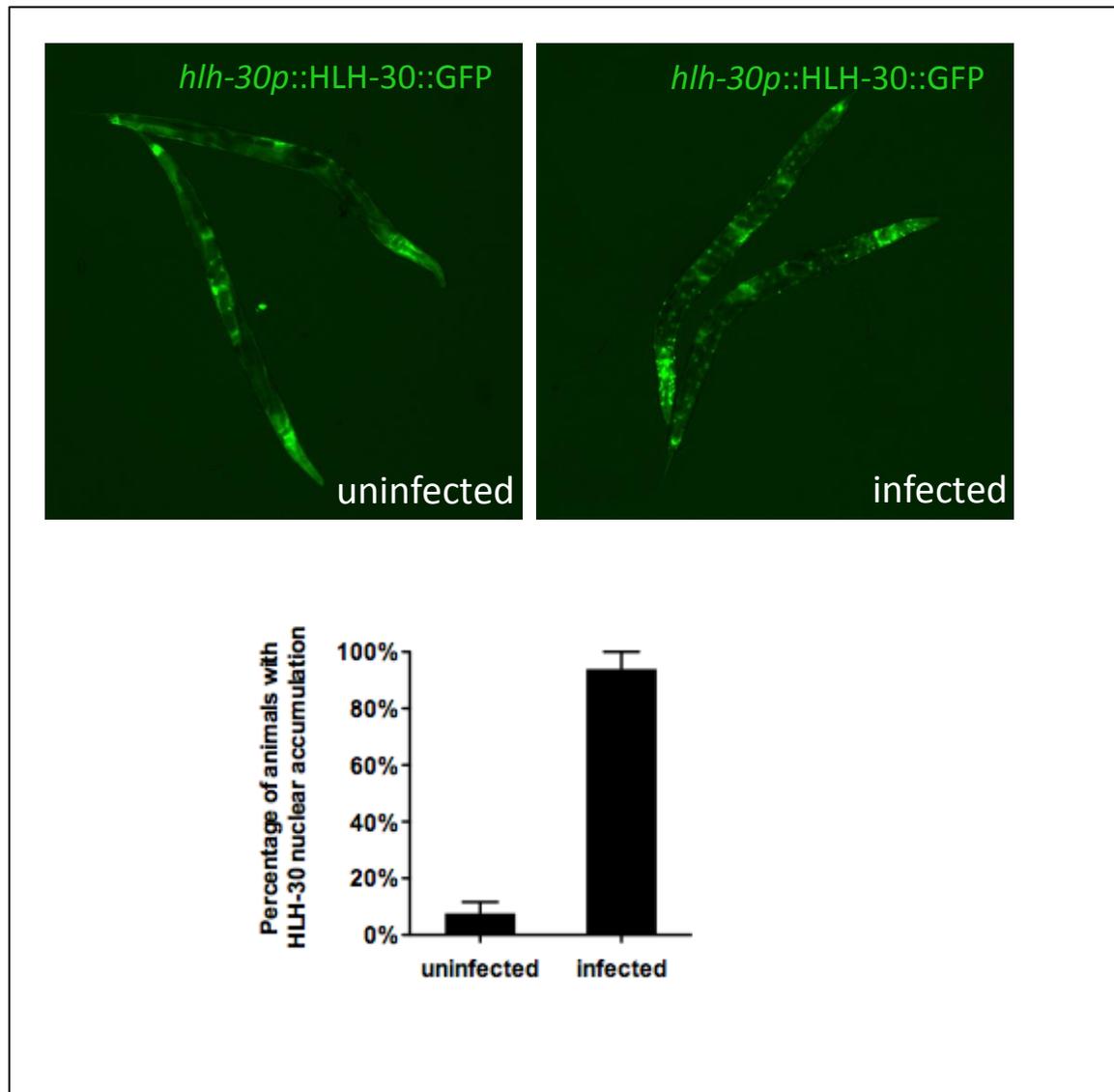
Transcriptional differences due to S. aureus infection

To gain insight on the role of HLH-30 in the transcriptional host response, we investigate which transcripts HLH-30 regulates in response to *S. aureus*. We used RNA-Seq to measure mRNA abundances of wild type and *hlh-30(-)* animals fed either non-pathogenic *E. coli* or *S. aureus* (i.e. 4 samples). Sequenced 42-bp reads from the two biological replicates in each condition were aligned to the *C. elegans* WS190 genome yielding an average of 12.7 million aligned reads for the 8 replicates (~21x coverage assuming a transcriptome of 25Mbp) (See Table 7 and Methods). The biological replicates are nearly identical as we observed little biological variance between them (r^2 values range from 0.94 to 0.97).

Table 7: Wild type and *hlh-30(-)* RNA-Sequencing reads for infected and uninfected samples

Replicate	mRNA Sample	Total Mapped Reads	Total Mapped Reads %
05_14_11	hlh-30_OP50	11,502,365	95.18%
	hlh-30_SA	13,433,320	95.45%
	N2_OP50	11,630,550	95.11%
	N2_SA	11,219,832	95.34%
05_26_11	hlh-30_OP50	11,690,588	95.42%
	hlh-30_SA	12,609,269	94.85%
	N2_OP50	13,341,375	95.45%
	N2_SA	11,383,914	95.28%

To avoid the variance overdispersion problem, we used DESeq (which uses a negative binomial distribution to model the expected variance) to determine significant *S. aureus* induced transcripts (SAITs) from our replicated RNA-Seq design (adjusted p<0.05) (Anders and Huber 2010). We implemented DESeq twice to determine the wild type SAITS and the mutant SAITs (i.e N2.SAITs and *hlh-30(-)*.SAITs).

Figure 21 shows the number of differentially expressed transcripts that are HLH-30 dependent (HLH-30.dep.SAITs), HLH-30 independent (HLH-30.indep.SAITs) and those that are differentially expressed even when lacking HLH-30 (HLH-30.comp.SAITs). We observe a significant overlap between genes with promoters that have highly occupied M-Box instances and the HLH-30.dep.SAITs set (p<4.6e-4). This enrichment is specific as the HLH-30.indep.SAITs (p<0.8484) and HLH-30.comp.SAITs sets (p<0.6318) are not significantly enriched for this motif. This specific enrichment again bolsters our argument that the Magma- discovered M-Box motif represents the preferred binding specificity for HLH-30.

Figure 21: A HLH-30 motif is specifically enriched in HLH-30-dep-SAITs

N2. SAITs (494)　　hlh-30(-).SAITs (247)

435　　59　　188

HLH-30-dep SAITs
(p<0.00046)

HLH-30-indep SAITs
(p<0.8484)

HLH-30-comp SAITs
(p<0.6318)
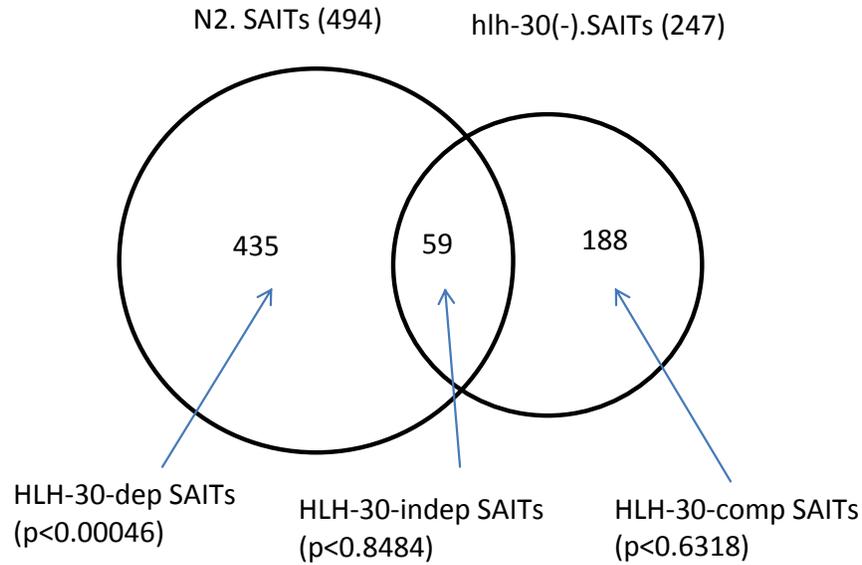
Validation of HLH-30 targets

To confirm the RNA-seq results, we measured the expression of 10 genes previously implicated with host defense in *C.elegans* using qRT-PCR (Irazoqui, Troemel et al. 2010). Six (6) of these were predicted to be HLH-30.dep.SAITs. Flavin-containing monooxygenase (*fmo-2*) is one of the predicted targets of HLH-30. A conserved exemplar site of the HLH-30-like M-Box motif lies just 195bp upstream of its translation start site. As seen in figure 22, fmo-2 is down-regulated by almost 8-fold when *hlh-30* is knocked out. All 10 genes are down-regulated by at least 2-fold in the mutant, and 8 of them down-regulated by at least 5-fold. This demonstrates that, in this set of targets,

HLH-30 is an important regulator involved in host-defense and, in the case of *fmo-2*, this regulation is likely direct.

Figure 22: HLH-30 targets are down-regulated in mutant animals versus wild type
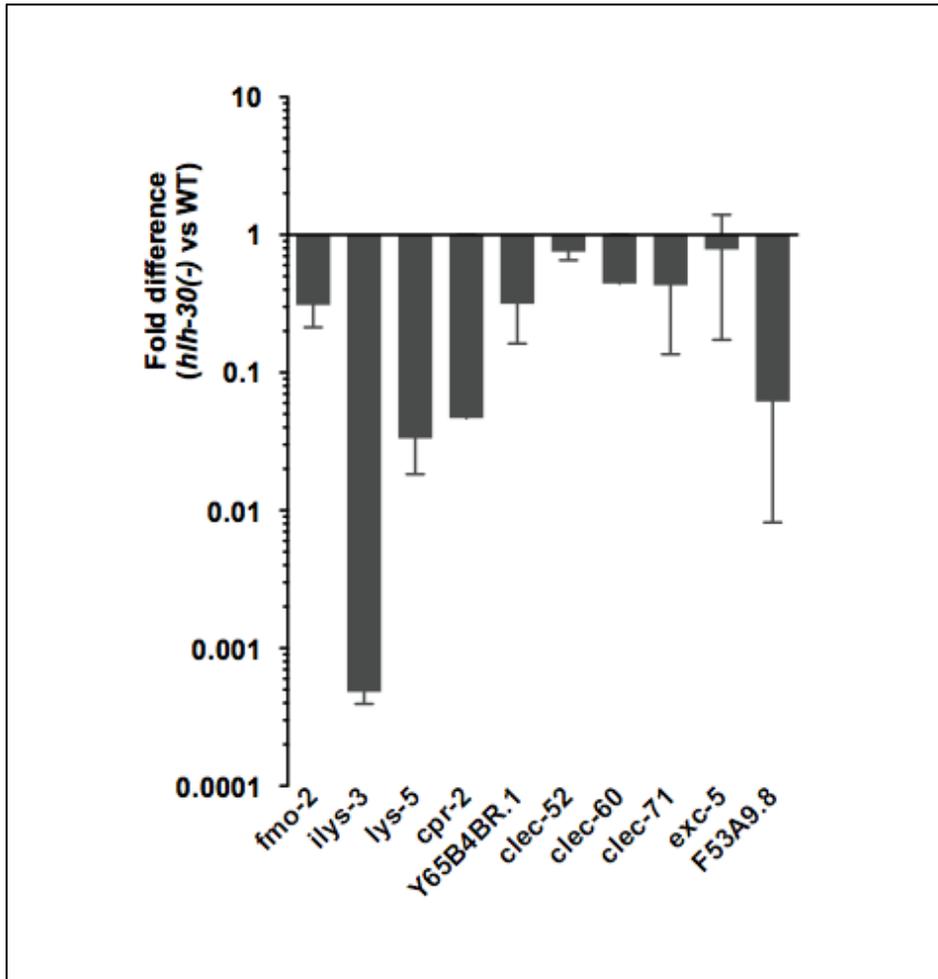
hlh-30(-) animals have significantly reduced lifespan due to infections

To determine the phenotype and impact of HLH-30 on host-defense and mortality, we performed _S. aureus_ killing assays comparing wild type animals to _hlh-30(-)_ mutants. Results show that _hlh-30(-)_ mutants animals display enhanced susceptibility to _S. aureus_ compared to wild type (Figure 23A). This results might be explained by a defective transcriptional response in _hlh-30(-)_ mutants. To determine if this response is specific to the Gram positive _S. aureus_ pathogen, we tested the susceptibility of _hlh-30(-)_ on the gram negative _Pseudomonas aeruginosa_. We performed killing assay using standard protocols (Powell and Ausubel 2008) and found that _hlh-30(-)_ animals are more susceptible to both pathogens (Figure 23 A & B) but not oxidative stress (Figure 23 C). In both killing assays, we see that _hlh-30(-)_ animals live about a day shorter than their wild type counterparts. This suggests that HLH-30 is a central regulator of the host response to multiple infections.

Figure 23: *hlh-30(-)* animals are more susceptible to infections
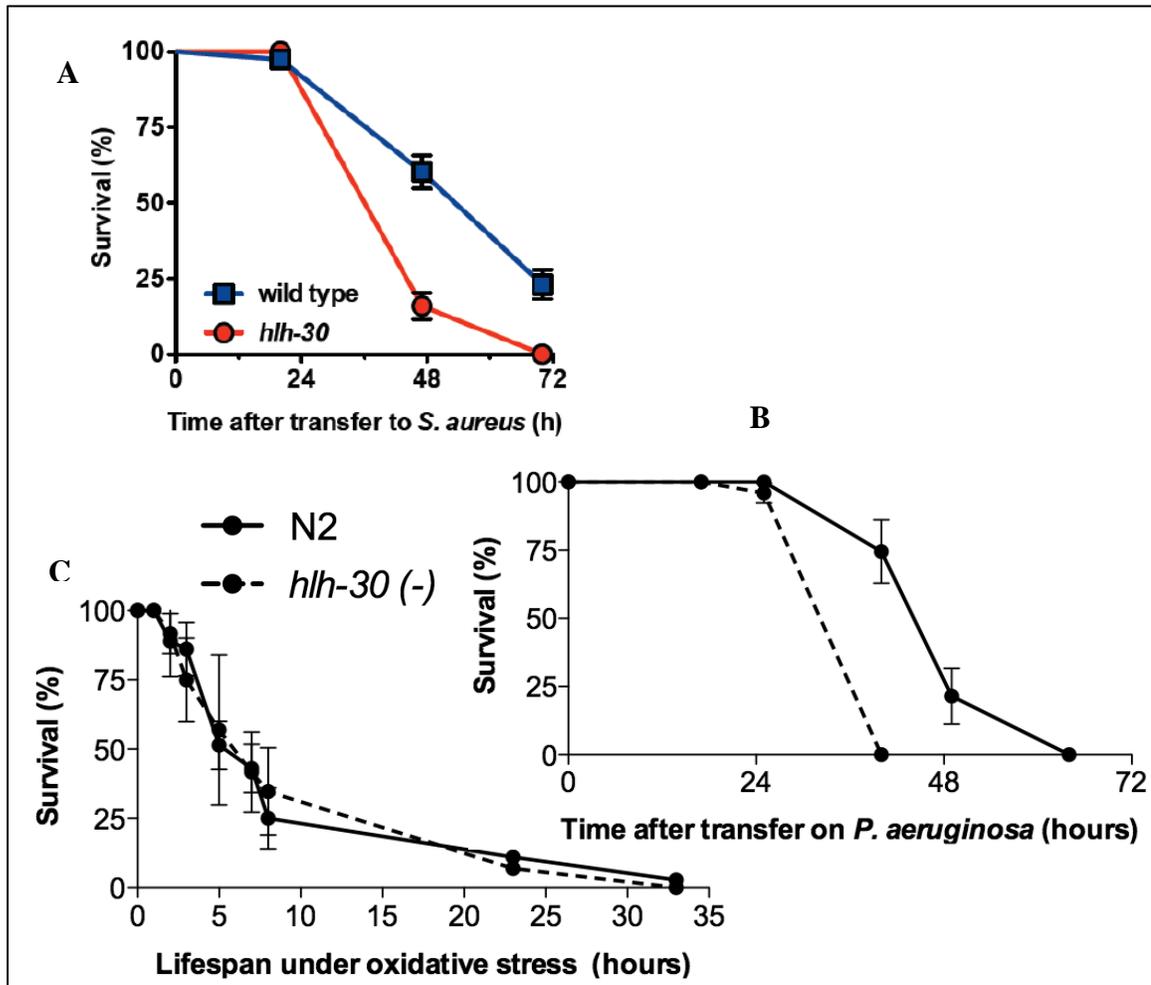
It is worth noting that *hlh-30(-)* animals have shorter lifespans than their wild type counterparts on non-pathogenic bacteria. As seen in Figure 24, *hlh-30(-)* animals typically live about 5 days shorter than wild type animals. Yet the stark infection phenotype difference we notice due to *hlh-30(-)* takes place relatively quickly within 2 days while the aging difference occurs after 5 days.

Figure 24: *hlh-30(-)* animals have shorter lifespans
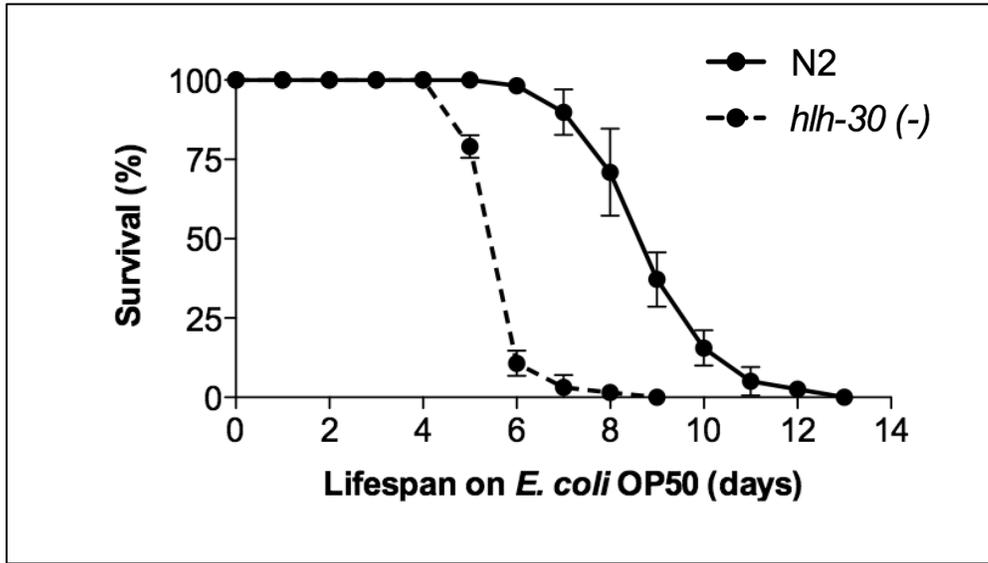
Methods and Materials

Strains

C. elegans strains used in this study are detailed in Table 8. Bacterial strains are detailed in Table 9. Transgenic animals where obtained by gonadal microinjection in wild type young gravid adults using pPRF4-rol-6 as a selection marker. *hlh-30p*::HLH-30::GFP fusion expression plasmid was obtained by LR recombinaison (Gateway system, invitrogen) using pDONR P4-P1R-*hlh-30p* (Open Biosystem), pDONR201-HLH-30 ORF (Vidal ORFeome) and pKA674 expression plasmid.

C. elegans Growth

C. elegans was grown on nematode-growth media (NGM) plates seeded with E. coli OP50 at 15–20°C according to standard procedures (Brenner 1974).

Cdc25 RNAi Knockdown

RNAi was carried out using bacterial feeding RNAi (Timmons, Court et al. 2001). L4 animals were incubated on Cdc25 RNAi bacteria for 48 h at 15°C before transfer to killing plates. Cdc25 RNAi clone was obtained from the Ahringer laboratory and confirmed by sequencing.

Killing Assays

S. aureus assays were performed as described in Sifri, Begun et al. (2003). *P. aeruginosa* killing assays were performed as described in Powell and Ausubel (2008). Briefly, NCTC8325 was grown overnight in tryptic soy broth (TSB, BD) with 10 μg/ml nalidixic

acid (Sigma). Ten µl of overnight cultures were seeded on 35 mm tryptic soy agar (TSA, BD) plates with 10µg/ml nalidixic acid, and incubated 4h at 37°C. PA14 was grown overnight in Luria Browth (LB, BD). Ten µl of overnight cultures were seeded on 35 mm slow killing plates, which contain modified NGM (0.35% peptone), incubated 24 h at 37°C then 24 h at 25°C, before adding 80–100 µg/ml 5-fluorodeoxyuridine (FUDR, Sigma), to prevent progeny from hatching. For *S. aureus*, a total of 25–35 Cdc25 RNAi treated animals were transferred to each of three replicate plates. For *P. aeruginosa*, a total of 25–35 L4 hermaphrodites were transferred to each of three replicate plates. Animals that died of bursting vulva or crawling off the agar were censored. Experiments were performed at least twice.

Lifespan assays

For lifespan assay, a total of 25–35 synchronized L4 hermaphrodites were transferred to each of three replicate plates per strain. For *E.coli* OP50 lifespan and heat-shock, NGM+OP50 plates were supplemented with 80 −100µg/ml FUDR. For lifespan under oxidative stress conditions, 6 synchronized L4 hermaphrodites were transferred in M9 supplemented with paraquat 100mM (Sigma) to each well of three replicate 12 well plates per strain. Experiments were performed at least twice.

Quantitative RT-PCR Analysis

Synchronized *C. elegans* animals were treated similar to the killing assays described above, where *S.aureus* infected samples were compared with parallel samples feeding on heat-killed E. coli OP50 on the same medium. Total RNA was extracted using TRI

Reagent (Molecular Research Center) and reverse transcribed by using the SuperScript III kit (Invitrogen). cDNA was subjected to qRT-PCR analysis as described (Irazoqui, Troemel et al. 2010). Primer sequences are detailed in Table 10. All values are normalized against the control gene *snb-1*, which did not vary under conditions being tested. Fold change was calculated by using the Pfaffl method (39). One-sample t tests were performed by using Graphpad Prism 4. A P value less than or equal to 0.05 was considered significant.

Table 8 : list of C. elegans strains

Strain	Relevant genotype	source
N2 Bristol	Wild type	CGC
VT1584	*hlh-30(tm1978)IV*	CGC
JIN1610	*jinEx[rol-6]*	This work
JIN1590	*jinEx[hlh-30p::HLH-30::GFP,rol-6]*	This work

Table 9 : list of bacterial strain

Bacterial species	Bacterial strains
Escherichia coli OP50	Ura- StrR
Staphylococcus aureus NCTC832	Wild type strain; rsbU mutant
Pseudomonas aeruginosa PA14	Pathogenic clinical isolate

Table 10: list of primer for qRT-PCR

Name	Sequence (5'-3')
snb-1 F	CCGGATAAGACCATCTTGACG
snb-1 R	GACGACTTCATCAACCTGAGC
clec-60 F	ACGGGCAAGTTATTGGAGAG
clec-60 R	ACACGGTATTGAATCCACGA
F53A9.8 F	GCGCTAAAACTCAACACCAA
F53A9.8 R	ATGTCCTTCATGGGAGTCGT

clec-52 F	ATGGAGGAGATTTGGCTTCA
clec-52 R	CCTGTCCAATCCTTGTCCTT
clec-71 F	CGGTATCGAGCAAGACTCAC
clec-71 R	GCATTGACGGCATATATTGG
exc-5 F	CCTGATGGATCAACAACAACA
exc-5 R	TAAGTCTCTTGGCGGGAGAA
lys-5 F	TCCCAGAATTTATCATTCATCG
lys-5 R	TGGCATTCTTGACATTTTGC
fmo-2 F	AAGCTGGAGACACGAGGATT
f mo-2 R	GGAGTTAAGCATAGCTTGAGGAA
ilys-3 F	GCGAATGATCTTAGCTGTGC
ilys-3 R	CCAGTTCCAGCACATTGACT
cpr-2 F	CAGAACGACCTACACCAACG
cpr-2 R	CGGTTCTTGGAACAGGGTAT
Y65B4BR.1 F	AAATGTGATCACTGCCATTCA
Y65B4BR.1 R	ATTCCGGTCATGGATACGAT

Uncovering HLH-30 from Expression Microarray

Irazoqui, Troemel et al. (2010) infected worms with *S. aureus,* RNA from replicates were extracted, labeled and hybridized to gene probes using the GPL200 Affymetrix *C. elegans* Genome Array (GSE21819). Assuming a normal distribution of the mean intensity fold-changes, we performed a Z-Score test and selected 688 genes with significantly increased log-fold changes (p<0.05).

For each of the 2309 Magma-discovered motifs (M) we defined an occupancy score similar to Granek and Clarke (2005) for each promoter (P) in the genome:

$$Occ\,(M,P) \approx \left[\sum_{i=1}^{L_p - L_M + 1} \prod_{j=1}^{L_M} M_j, nt_{(P_{i+j-1})} \right]$$

In the above equation, M is a motif matrix containing the relative frequency of each base with respect to the consensus nucleotide. This is also known as a Position-Specific Affinity Matrix or PSAM (Foat, Houshmandi et al. 2005). The parameters L_p and L_M are the respective lengths of the promoter and the motif. Using this score, a sum of 1 or better means the accrued sites for this motif along the promoter is as good as a consensus site. For each motif, we collected all promoters with occupancy scores equivalent to at least one consensus sequence and performed right-tailed fisher exact tests with this list and the 688 genes to uncover enriched motifs. We report the enrichment p-value for each motif after correcting for multiple hypothesis testing.

Method 2: Discovering differentially expressed transcripts using RNA-Seq

After sequencing, raw 42-bp reads were aligned to the WS190 assembly of C. elegans using TopHat (Trapnell, Pachter et al. 2009). The abundance of transcripts were estimated and normalized as FPKMs using Cufflinks (Mortazavi, Williams et al. 2008; Trapnell, Williams et al. 2010). We used DESeq to determine differentially expressed transcripts between the uninfected and infected populations, given the two biological replicates, and only retained transcripts that were significant at an adjust p-value of less than 0.05.

Discussion and Conclusion

In this chapter we show that a Magma-discovered M-Box motif is enriched in the promoters of up-regulated genes upon *S. aureus* induction. This motif is similar to that discovered for HLH-30 using *in vitro* protein binding microarrays (Grove, De Masi et al. 2009). Using qRT-PCR we show that hlh-30 is up-regulated 2-fold upon *S. aureus* infection. Furthermore, in about 30 minutes much of the already translated HLH-30 rushes into the nucleus in response to infection (~95% of animals showed a punctate fluorescent pattern). This increased transcription and available proteins executes a differential transcriptional program in which 435 transcripts show significant specific HLH-30-dependent *S. aureus* induction (SAITs). These SAITs are enriched for metabolic processes and their regulation by HLH-30 is critical for surviving the infection. In fact we show that HLH-30 is critical for surviving both the gram positive *S. aureus* and the gram negative *P. auereginosa* suggesting it is a central regulator in the hosts' response to pathogenic bacteria. *hlh-30(-)* animals tend to live about 1 day shorter than their wild type counterparts due to infection. Interestingly, *hlh-30* seems to also impact aging as these mutant animals also live much shorter than wild type animals and aging-related genes are overrepresented in the HLH-30.dep.SAITs. Yet this aging phenotype occurs much later after the infection phenotype. The molecular pathways relating aging and the ability to fight infections are not well defined but HLH-30 seems to play a central role in both and may be a productive target to further those studies.

Finally, HLH-30 is homologous to the mammalian factor Microphthalmia factor (MiTF). Previous studies have implicated MiTF in various stresses (Saha, Singh et al. 2006; Liu, Fu et al. 2009) but not involved in pathogenic infections. Further studies will be required

to see if HLH-30's role in regulating differential transcription due to infections is

conserved in mammals via MiTF.

Chapter 6: Conclusions and Future Directions

This dissertation had two major aims: (1) To catalog all *cis*-regulatory elements within the intergenic and intronic regions surrounding every gene in *C. elegans* (i.e. the regulome) and (2) to determine *cis*-regulatory elements associated with expression under specific conditions. Applying PhyloNet and CERMOD was a significant initial step to achieving the first aim. We found motifs that matched several known transcription factor binding sites and other known regulatory elements. Additionally our module predictions overlapped most of the known modules. Yet these initial results had a lot of redundant motifs and the approach was not efficiently scalable to the entire *C. elegans* regulome. Magma (Multiple Aligner of Genomic Multiple Alignments) overcomes these shortcomings by: (1) utilizing more efficient HSP clustering methods that offer strong performance and quality guarantees. It uses interval clique finding to ensure maximality and exhaustive enumeration of clusters, with better scaling than general clique finding. (2) Magma uses an enumerative algorithm to convert HSP clusters to motifs that requires $\Theta(m^2 n)$, where n is the number of HSPs in the cluster and m is the interval coverage of the largest cluster. However efficient use of lookup tables makes the quadratic cost small in practice. (3) Magma uses a fast greedy set-cover solution to achieve a log(n) approximation to the motif redundancy problem. These differences allowed Magma to predict about 2300 motifs and 110,000 CRMs in the intergenic and intronic regions of about 99% of all protein-coding genes in *C. elegans*. Additionally we show that the approach tractably scales to higher order organisms with larger regulomes.

Although I believe this is a milestone in motif-finding research, there are still other pertinent information that has been challenging for motif-finders to incorporate, but will mark the next-generation of these tools. Due to lack of TF concentration and chromatin accessibility information, most motif-finders predict more putative transcription factor binding sites than are observed to be bound by TFs (by ChIP-CHIP, ChIP-Seq and other methods). Hence the ability to incorporate contextual cell-state information such as the chromatin structure and other epigenetic marks will help make much more specific and informed motif predictions that reflect the functional state of the cell. The modENCODE project is currently pioneering the measurement of these epigenetic marks in different cells throughout *C.elegans* (Gerstein, Lu et al. 2010). As more of these kinds of information become available, motif-finders will be best served to incorporate their information into their models, similar to how conservation information has been incorporated.

To achieve the second aim of this thesis, we show functional enrichments of the predicted motifs in various expression datasets. The implication is that the motifs represent the binding specificities of TFs involved in the particular expression being measured. Using the GEO Omnibus database of expression microarray results we make hundreds of these predictions. In some cases we implicate known elements in novel functions (such as the possible role of GATA elements in response to Cadmium exposure) and in other times we see new regions of regulation (such as the enrichment of GATA elements in the introns of genes expressed in the uterus).

Specifically we predict and validate a novel role for HLH-30 in the host response to infection. *S. aureus and P. aeruginosa* are among the most common bacterial

infections in humans and cause many life-threatening symptoms. Lately, the increasing

prevalence of Methicillin-resistant *S. aureus* (MRSA) is motivating a lot of studies to find

novel pathways to combat the pathogen. HLH-30 is homologous to MiTF, a mammalian

transcription factor that had previously been associated to some stress conditions but not

infections. Inspecting the promoters of up-regulated genes in a *S. aureus* infection

microarray, we discover an enrichment of a Magma-discovered M-Box motif. This motif

was subsequently shown to be a binding preference for HLH-30. Results describe how

the *hlh-30* gene is up-regulated, and has increased nuclear localization in response to *S.*

aureus. When *hlh-30* is knocked out to create *hlh-30(-)* animals its predicted targets,

which include previously implicated genes in host-defense, are also down-regulated

under *S. aureus*. The predicted HLH-30-like M-Box motif is specifically enriched in this

set of HLH-30-dependent *S. aureus* induced transcripts (SAITs). Finally, compared to

their wild type counterparts, *hlh-30(-)* animals are not able to respond as well to

infections and are more susceptible to their fatal effects. This infection-caused fatality

occurs within 2-3 days, much quicker than the aging complications caused by knocking

out the *hlh-30* gene.

It will beneficial to know if MiTF (a human homolog of HLH-30) is a similar

regulator of the host-defense to these pathogens. MiTF has several well defined targets,

do these also show a MiTF-dependent infection induced up-regulation? How disruptive is

the knockout of MiTF on the host's ability to fight infection? In the case *C. elegans,* it

was extremely fatal as its intestinal epithelial cells are non-renewable. Fortunately, this is

not the case for humans. Can transgenically-increased levels of MiTF decrease

susceptibility to these pathogens? Answers to these questions will provide valuable

insight of the conserved role of HLH-30/MiTF as a central regulator in the host's defense

to pathogens and, depending on their results, could eventually present a new potential

target for anti-microbial drugs.

Chapter 7: References

Aballay, A. and F. M. Ausubel (2002). "Caenorhabditis elegans as a host for the study of host-pathogen interactions." Curr Opin Microbiol **5**(1): 97-101.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." Genome Biol **11**(10): R106.

Antebi, A. (2007). "Genetics of aging in Caenorhabditis elegans." PLoS Genet **3**(9): 1565-1571.

Anyanful, A., Y. Sakube, et al. (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**(3): 525-537.

Ao, W., J. Gaudet, et al. (2004). "Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR." Science **305**(5691): 1743-1746.

Arnone, M. I. and E. H. Davidson (1997). "The hardwiring of development: organization and function of genomic regulatory systems." Development **124**(10): 1851-1864.

Bailey, T. L., N. Williams, et al. (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." Nucleic Acids Res **34**(Web Server issue): W369-373.

Barrasa, M. I., P. Vaglio, et al. (2007). "EDGEdb: a transcription factor-DNA interaction database for the analysis of C. elegans differential gene expression." BMC Genomics **8**: 21.

Barrett, T., D. B. Troup, et al. (2009). "NCBI GEO: archive for high-throughput functional genomic data." Nucleic Acids Res **37**(Database issue): D885-890.

Baugh, L. R., J. Demodena, et al. (2009). "RNA Pol II accumulates at promoters of growth genes during developmental arrest." Science **324**(5923): 92-94.

Beer, M. A. and S. Tavazoie (2004). "Predicting gene expression from sequence." Cell **117**(2): 185-198.

Berg, O. G. and P. H. von Hippel (1987). "Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters." J Mol Biol **193**(4): 723-750.

Blanchette, M., A. R. Bataille, et al. (2006). "Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression." <u>Genome Res</u> **16**(5): 656-668.

Blumenthal, T. and K. Steward (1997). RNA Processing and Gene Structure. <u>C. ELEGANS</u>. D. L. Riddle, T. Blumenthal, B. J. Meyer and J. R. Priess, Cold Spring Harbor Laboratory Press.

Bolouri, H. and E. H. Davidson (2002). "Modeling DNA sequence-based cis-regulatory gene networks." <u>Dev Biol</u> **246**(1): 2-13.

Boucher, H. W. and G. R. Corey (2008). "Epidemiology of methicillin-resistant Staphylococcus aureus." <u>Clin Infect Dis</u> **46 Suppl 5**: S344-349.

Boulin, T., J. F. Etchberger, et al. (2006). "Reporter gene fusions." <u>WormBook</u>: 1-23.

Boyerinas, B., S. M. Park, et al. (2010). "The role of let-7 in cell differentiation and cancer." <u>Endocr Relat Cancer</u> **17**(1): F19-36.

Brenner, S. (1974). "The genetics of Caenorhabditis elegans." <u>Genetics</u> **77**(1): 71-94.

Brodigan, T. M., J. Liu, et al. (2003). "Cyclin E expression during development in Caenorhabditis elegans." <u>Dev Biol</u> **254**(1): 102-115.

Bussemaker, H. J., B. C. Foat, et al. (2007). "Predictive modeling of genome-wide mRNA expression: from modules to molecules." <u>Annu Rev Biophys Biomol Struct</u> **36**: 329-347.

Bussemaker, H. J., H. Li, et al. (2001). "Regulatory element detection using correlation with expression." <u>Nat Genet</u> **27**(2): 167-171.

Casamassimi, A. and C. Napoli (2007). "Mediator complexes and eukaryotic transcription regulation: an overview." <u>Biochimie</u> **89**(12): 1439-1446.

Celniker, S. E., L. A. Dillon, et al. (2009). "Unlocking the secrets of the genome." <u>Nature</u> **459**(7249): 927-930.

Chang, L. W., R. Nagarajan, et al. (2006). "A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles." <u>Genome Res</u> **16**(3): 405-413.

Chikina, M. D., C. Huttenhower, et al. (2009). "Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans." <u>PLoS Comput Biol</u> **5**(6): e1000417.

Cho, J. H., S. H. Eom, et al. (1999). "Analysis of calsequestrin gene expression using green fluorescent protein in Caenorhabditis elegans." Mol Cells **9**(2): 230-234.

Chvatal, V. (1979). "A greedy heuristic for the set-covering problem." Mathematics of Operations Research **4**(3): 233-235.

Consortium., *C. e.* S. (1998). "Genome sequence of the nematode *C. elegans*: a platform for investigating biology." Science **282**(5396): 2012-2018.

Couillault, C. and J. J. Ewbank (2002). "Diverse bacteria are pathogens of Caenorhabditis elegans." Infect Immun **70**(8): 4705-4707.

Cui, M. and M. Han (2003). "Cis regulatory requirements for vulval cell-specific expression of the Caenorhabditis elegans fibroblast growth factor gene egl-17." Dev Biol **257**(1): 104-116.

Culetto, E., D. Combes, et al. (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**(5): 951-966.

Cuny, C., A. Friedrich, et al. (2010). "Emergence of methicillin-resistant Staphylococcus aureus (MRSA) in different animal species." Int J Med Microbiol **300**(2-3): 109-117.

Deplancke, B., A. Mukhopadhyay, et al. (2006). "A gene-centered C. elegans protein-DNA interaction network." Cell **125**(6): 1193-1205.

Diep, B. A. and M. Otto (2008). "The role of virulence determinants in community-associated MRSA pathogenesis." Trends Microbiol **16**(8): 361-369.

Dupuy, D., Q. R. Li, et al. (2004). "A first version of the *Caenorhabditis elegans* Promoterome." Genome Res **14**(10B): 2169-2175.

Eastman, C., H. R. Horvitz, et al. (1999). "Coordinated transcriptional regulation of the unc-25 glutamic acid decarboxylase and the unc-47 GABA vesicular transporter by the Caenorhabditis elegans UNC-30 homeodomain protein." J Neurosci **19**(15): 6225-6234.

Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci U S A **95**(25): 14863-14868.

Elemento, O., N. Slonim, et al. (2007). "A universal framework for regulatory element discovery across all genomes and data types." Mol Cell **28**(2): 337-350.

Esquela-Kerscher, A., S. M. Johnson, et al. (2005). "Post-embryonic expression of C. elegans microRNAs belonging to the lin-4 and let-7 families in the hypodermis and the reproductive system." Dev Dyn **234**(4): 868-877.

Etchberger, J. F., A. Lorch, et al. (2007). "The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron." Genes Dev **21**(13): 1653-1674.

Farnham, P. J. (2009). "Insights from genomic profiling of transcription factors." Nat Rev Genet **10**(9): 605-616.

Ferretti, V., C. Poitras, et al. (2007). "PReMod: a database of genome-wide mammalian cis-regulatory module predictions." Nucleic Acids Res **35**(Database issue): D122-126.

Fire, A., S. W. Harrison, et al. (1990). "A modular set of lacZ fusion vectors for studying gene expression in Caenorhabditis elegans." Gene **93**(2): 189-198.

Foat, B. C., S. S. Houshmandi, et al. (2005). "Profiling condition-specific, genome-wide regulation of mRNA stability in yeast." Proc Natl Acad Sci U S A **102**(49): 17675-17680.

Foat, B. C., A. V. Morozov, et al. (2006). "Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE." Bioinformatics **22**(14): e141-149.

Frazer, K. A., L. Elnitski, et al. (2003). "Cross-species sequence comparisons: a review of methods and available resources." Genome Res **13**(1): 1-12.

Gaudet, J. and S. E. Mango (2002). "Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4." Science **295**(5556): 821-825.

Gaudet, J., S. Muttumu, et al. (2004). "Whole-genome analysis of temporal gene expression during foregut development." PLoS Biol **2**(11): e352.

Gaudet, J., I. VanderElst, et al. (1996). "Post-transcriptional regulation of sex determination in Caenorhabditis elegans: widespread expression of the sex-determining gene fem-1 in both sexes." Mol Biol Cell **7**(7): 1107-1121.

Gelfand, M. S. (1999). "Recognition of regulatory sites by genomic comparison." Res Microbiol **150**(9-10): 755-771.

Gerstein, M. B., Z. J. Lu, et al. (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." Science **330**(6012): 1775-1787.

Gerstein, M. B., Z. J. Lu, et al. (2011). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." Science **330**(6012): 1775-1787.

Gertz, J., E. D. Siggia, et al. (2009). "Analysis of combinatorial cis-regulation in synthetic and genomic promoters." Nature **457**(7226): 215-218.

Gilleard, J. S., J. D. Barry, et al. (1997). "cis regulatory requirements for hypodermal cell-specific expression of the Caenorhabditis elegans cuticle collagen gene dpy-7." Mol Cell Biol **17**(4): 2301-2311.

Gordon, R. J. and F. D. Lowy (2008). "Pathogenesis of methicillin-resistant Staphylococcus aureus infection." Clin Infect Dis **46 Suppl 5**: S350-359.

Graham, P. L., 3rd, S. X. Lin, et al. (2006). "A U.S. population-based survey of Staphylococcus aureus colonization." Ann Intern Med **144**(5): 318-325.

Granek, J. A. and N. D. Clarke (2005). "Explicit equilibrium modeling of transcription-factor binding and gene regulation." Genome Biol **6**(10): R87.

Griffith, O. L., S. B. Montgomery, et al. (2008). "ORegAnno: an open-access community-driven resource for regulatory annotation." Nucleic Acids Res **36**(Database issue): D107-113.

Grove, C. A., F. De Masi, et al. (2009). "A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors." Cell **138**(2): 314-327.

GuhaThakurta, D. (2006). "Computational identification of transcriptional regulatory elements in DNA sequence." Nucleic Acids Res **34**(12): 3585-3598.

GuhaThakurta, D., L. Palomar, et al. (2002). "Identification of a novel cis-regulatory element involved in the heat shock response in Caenorhabditis elegans using microarray gene expression and computational methods." Genome Res **12**(5): 701-712.

GuhaThakurta, D., L. A. Schriefer, et al. (2004). "Novel transcription regulatory elements in Caenorhabditis elegans muscle genes." Genome Res **14**(12): 2457-2468.

Gupta, B. P. and P. W. Sternberg (2002). "Tissue-specific regulation of the LIM homeobox gene lin-11 during development of the Caenorhabditis elegans egg-laying system." Dev Biol **247**(1): 102-115.

Gupta, U. I., D. T. Lee, et al. (1982). "Efficient Algorithms for Interval Graphs and Circular Arc-Graphs." <u>Networks</u> **12**: 459-467.

Harfe, B. D. and A. Fire (1998). "Muscle and nerve-specific regulation of a novel NK-2 class homeodomain factor in Caenorhabditis elegans." <u>Development</u> **125**(3): 421-429.

Harfe, B. D., A. Vaz Gomes, et al. (1998). "Analysis of a Caenorhabditis elegans Twist homolog identifies conserved and divergent aspects of mesodermal patterning." <u>Genes Dev</u> **12**(16): 2623-2635.

Hebbar, P. B. and T. K. Archer (2003). "Nuclear factor 1 is required for both hormone-dependent chromatin remodeling and transcriptional activation of the mouse mammary tumor virus promoter." <u>Mol Cell Biol</u> **23**(3): 887-898.

Henderson, S. T. and T. E. Johnson (2001). "daf-16 integrates developmental and environmental inputs to mediate aging in the nematode Caenorhabditis elegans." <u>Curr Biol</u> **11**(24): 1975-1980.

Hertz, G. Z. and G. D. Stormo (1999). "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences." <u>Bioinformatics</u> **15**(7-8): 563-577.

Hobert, O. (2008). "Regulatory logic of neuronal diversity: terminal selector genes and selector motifs." <u>Proc Natl Acad Sci U S A</u> **105**(51): 20067-20071.

Hobert, O. (2010). "Neurogenesis in the nematode Caenorhabditis elegans." <u>WormBook</u>: 1-24.

Hodgkin, J., P. E. Kuwabara, et al. (2000). "A novel bacterial pathogen, Microbacterium nematophilum, induces morphological change in the nematode C. elegans." <u>Curr Biol</u> **10**(24): 1615-1618.

Hunt-Newbury, R., R. Viveiros, et al. (2007). "High-throughput in vivo analysis of gene expression in Caenorhabditis elegans." <u>PLoS Biol</u> **5**(9): e237.

Hwang, B. J. and P. W. Sternberg (2004). "A cell-specific enhancer that specifies lin-3 expression in the C. elegans anchor cell for vulval development." <u>Development</u> **131**(1): 143-151.

Hwang, S. B. and J. Lee (2003). "Neuron cell type-specific SNAP-25 expression driven by multiple regulatory elements in the nematode Caenorhabditis elegans." J Mol Biol **333**(2): 237-247.

Ihuegbu, N., G. D. Stormo, et al. (2011). "Fast, sensitive discovery of conserved genome-wide motifs." J Comput Biol **Submitted**.

Irazoqui, J. E., A. Ng, et al. (2008). "Role for beta-catenin and HOX transcription factors in Caenorhabditis elegans and mammalian host epithelial-pathogen interactions." Proc Natl Acad Sci U S A **105**(45): 17469-17474.

Irazoqui, J. E., E. R. Troemel, et al. (2010). "Distinct pathogenesis and host responses during infection of C. elegans by P. aeruginosa and S. aureus." PLoS Pathog **6**: e1000982.

Irazoqui, J. E., J. M. Urbach, et al. (2010). "Evolution of host innate defence: insights from Caenorhabditis elegans and primitive invertebrates." Nat Rev Immunol **10**(1): 47-58.

Iser, W. B., M. A. Wilson, et al. (2011). "Co-regulation of the DAF-16 target gene, cyp-35B1/dod-13, by HSF-1 in C. elegans dauer larvae and daf-2 insulin pathway mutants." PLoS One **6**(3): e17369.

Jans, J., J. M. Gladden, et al. (2009). "A condensin-like dosage compensation complex acts at a distance to control expression throughout the genome." Genes Dev **23**(5): 602-618.

Jensen, S. T., L. Shen, et al. (2005). "Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes." Bioinformatics **21**(20): 3832-3839.

Kagawa, H., K. Sugimoto, et al. (1995). "Genome structure, mapping and expression of the tropomyosin gene tmy-1 of Caenorhabditis elegans." J Mol Biol **251**(5): 603-613.

Kamath, R. S. and J. Ahringer (2003). "Genome-wide RNAi screening in Caenorhabditis elegans." Methods **30**(4): 313-321.

Karp, R. M. (1972). "Reducibility Among Combinatorial Problems." Complexity of Computer Computations: 85-103.

Keles, S., M. van der Laan, et al. (2002). "Identification of regulatory elements using a feature selection method." Bioinformatics **18**(9): 1167-1175.

Kielbasa, S. M., D. Gonze, et al. (2005). "Measuring similarities between transcription factor binding sites." <u>BMC Bioinformatics</u> **6**: 237.

Kimble, J. and D. Hirsh (1979). "The postembryonic cell lineages of the hermaphrodite and male gonads in Caenorhabditis elegans." <u>Dev Biol</u> **70**(2): 396-417.

King, D. C., J. Taylor, et al. (2005). "Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences." <u>Genome Res</u> **15**(8): 1051-1060.

Kipreos, E. T. (2005). "C. elegans cell cycles: invariance and stem cell divisions." <u>Nat Rev Mol Cell Biol</u> **6**(10): 766-776.

Kirchhamer, C. V., C. H. Yuh, et al. (1996). "Modular cis-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples." <u>Proc Natl Acad Sci U S A</u> **93**(18): 9322-9328.

Kirouac, M. and P. W. Sternberg (2003). "cis-Regulatory control of three cell fate-specific genes in vulval organogenesis of Caenorhabditis elegans and C. briggsae." <u>Dev Biol</u> **257**(1): 85-103.

Kolbe, D., J. Taylor, et al. (2004). "Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat." <u>Genome Res</u> **14**(4): 700-707.

Koo, J., Y. Kim, et al. (2007). "A GUS/luciferase fusion reporter for plant gene trapping and for assay of promoter activity with luciferin-dependent control of the reporter protein stability." <u>Plant Cell Physiol</u> **48**(8): 1121-1131.

Krause, M., S. W. Harrison, et al. (1994). "Elements regulating cell- and stage-specific expression of the C. elegans MyoD family homolog hlh-1." <u>Dev Biol</u> **166**(1): 133-148.

Kuchenthal, C. A., W. Chen, et al. (2001). "Multiple enhancers contribute to expression of the NK-2 homeobox gene ceh-22 in C. elegans pharyngeal muscle." <u>Genesis</u> **31**(4): 156-166.

Kurz, C. L. and J. J. Ewbank (2000). "Caenorhabditis elegans for the study of host-pathogen interactions." <u>Trends Microbiol</u> **8**(3): 142-144.

Kurz, C. L. and J. J. Ewbank (2003). "Caenorhabditis elegans: an emerging genetic model for the study of innate immunity." <u>Nat Rev Genet</u> **4**(5): 380-390.

Lall, S., C. C. Friedman, et al. (2004). "Contribution of trans-splicing, 5' -leader length, cap-poly(A) synergism, and initiation factors to nematode translation in an Ascaris suum embryo cell-free system." J Biol Chem **279**(44): 45573-45585.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lawrence, C. E., S. F. Altschul, et al. (1993). "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment." Science **262**(5131): 208-214.

Lazakovitch, E., J. M. Kalb, et al. (2008). "Lifespan extension and increased pumping rate accompany pharyngeal muscle-specific expression of nfi-1 in C. elegans." Dev Dyn **237**(8): 2100-2107.

Lazakovitch, E., J. M. Kalb, et al. (2005). "nfi-I affects behavior and life-span in C. elegans but is not essential for DNA replication or survival." BMC Dev Biol **5**: 24.

Lehner, B., C. Crombie, et al. (2006). "Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways." Nat Genet **38**(8): 896-903.

Lehner, B., J. Tischler, et al. (2006). "RNAi screens in Caenorhabditis elegans in a 96-well liquid format and their application to the systematic identification of genetic interactions." Nat Protoc **1**(3): 1617-1620.

Li, X. Y., S. Thomas, et al. (2011). "The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding." Genome Biol **12**(4): R34.

Liu, F., Y. Fu, et al. (2009). "MiTF regulates cellular response to reactive oxygen species through transcriptional regulation of APE-1/Ref-1." J Invest Dermatol **129**(2): 422-431.

Liu, X., F. Long, et al. (2009). "Analysis of cell fate from single-cell gene expression profiles in C. elegans." Cell **139**(3): 623-633.

Luscombe, N. M., S. E. Austin, et al. (2000). "An overview of the structures of protein-DNA complexes." Genome Biol **1**(1): REVIEWS001.

MacMorris, M., S. Broverman, et al. (1992). "Regulation of vitellogenin gene expression in transgenic Caenorhabditis elegans: short sequences required for activation of the vit-2 promoter." Mol Cell Biol **12**(4): 1652-1662.

Maduro, M. F. (2006). "Endomesoderm specification in Caenorhabditis elegans and other nematodes." Bioessays **28**(10): 1010-1022.

Maduro, M. F., G. Broitman-Maduro, et al. (2007). "Maternal deployment of the embryonic SKN-1-->MED-1,2 cell specification pathway in C. elegans." Dev Biol **301**(2): 590-601.

Maduro, M. F. and J. H. Rothman (2002). "Making worm guts: the gene regulatory network of the Caenorhabditis elegans endoderm." Dev Biol **246**(1): 68-85.

Mahony, S., P. E. Auron, et al. (2007). "DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies." PLoS Comput Biol **3**(3): e61.

Mango, S. E. (2007). "The C. elegans pharynx: a model for organogenesis." WormBook: 1-26.

Mango, S. E. (2009). "The molecular basis of organ formation: insights from the C. elegans foregut." Annu Rev Cell Dev Biol **25**: 597-628.

Marchal, K., S. De Keersmaecker, et al. (2004). "In silico identification and experimental validation of PmrAB targets in Salmonella typhimurium by regulatory motif detection." Genome Biol **5**(2): R9.

Matys, V., E. Fricke, et al. (2003). "TRANSFAC: transcriptional regulation, from patterns to profiles." Nucleic Acids Res **31**(1): 374-378.

McCue, L., W. Thompson, et al. (2001). "Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes." Nucleic Acids Res **29**(3): 774-782.

McElwee, J. J., E. Schuster, et al. (2004). "Shared transcriptional signature in Caenorhabditis elegans Dauer larvae and long-lived daf-2 mutants implicates detoxification system in longevity assurance." J Biol Chem **279**(43): 44533-44543.

McGhee, J. D., T. Fukushige, et al. (2009). "ELT-2 is the predominant transcription factor controlling differentiation and function of the C. elegans intestine, from embryo to adult." Dev Biol **327**(2): 551-565.

McGhee, J. D., M. C. Sleumer, et al. (2007). "The ELT-2 GATA-factor and the global regulation of transcription in the C. elegans intestine." Dev Biol **302**(2): 627-645.

McGuire, A. M., J. D. Hughes, et al. (2000). "Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes." Genome Res **10**(6): 744-757.

Moilanen, L. H., T. Fukushige, et al. (1999). "Regulation of metallothionein gene transcription. Identification of upstream regulatory elements and transcription factors responsible for cell-specific expression of the metallothionein genes from Caenorhabditis elegans." J Biol Chem **274**(42): 29655-29665.

Monsieurs, P., G. Thijs, et al. (2006). "More robust detection of motifs in coexpressed genes by using phylogenetic information." BMC Bioinformatics **7**: 160.

Montgomery, S. B., O. L. Griffith, et al. (2006). "ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation." Bioinformatics **22**(5): 637-640.

Mortazavi, A., B. A. Williams, et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-628.

Nachman, I., A. Regev, et al. (2004). "Inferring quantitative models of regulatory networks from expression data." Bioinformatics **20 Suppl 1**: i248-256.

Nam, S., Y. H. Jin, et al. (2002). "Expression pattern, regulation, and biological role of runt domain transcription factor, run, in Caenorhabditis elegans." Mol Cell Biol **22**(2): 547-554.

Narlikar, L. and A. J. Hartemink (2006). "Sequence features of DNA binding sites reveal structural class of associated transcription factor." Bioinformatics **22**(2): 157-163.

Natarajan, L., B. M. Jackson, et al. (2004). "Identification of evolutionarily conserved promoter elements and amino acids required for function of the C. elegans beta-catenin homolog BAR-1." Dev Biol **272**(2): 536-557.

Neves, A., K. English, et al. (2007). "Notch-GATA synergy promotes endoderm-specific expression of ref-1 in C. elegans." Development **134**(24): 4459-4468.

Niwa, R. and K. Hada (2010). "Identification of a spatio-temporal enhancer element for the Alzheimer's amyloid precursor protein-like-1 gene in the nematode Caenorhabditis elegans." Biosci Biotechnol Biochem **74**(12): 2497-2500.

Nizet, V. (2007). "Understanding how leading bacterial pathogens subvert innate immunity to reveal novel therapeutic targets." J Allergy Clin Immunol **120**(1): 13-22.

Okkema, P. G., S. W. Harrison, et al. (1993). "Sequence requirements for myosin gene expression and regulation in Caenorhabditis elegans." Genetics **135**(2): 385-404.

Oommen, K. S. and A. P. Newman (2007). "Co-regulation by Notch and Fos is required for cell fate specification of intermediate precursors during C. elegans uterine development." Development **134**(22): 3999-4009.

Panina, E. M., A. A. Mironov, et al. (2001). "Comparative analysis of FUR regulons in gamma-proteobacteria." Nucleic Acids Res **29**(24): 5195-5206.

Panina, E. M., A. G. Vitreschak, et al. (2003). "Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria." FEMS Microbiol Lett **222**(2): 211-220.

Pauli, F., Y. Liu, et al. (2006). "Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in C. elegans." Development **133**(2): 287-295.

Pilpel, Y., P. Sudarsanam, et al. (2001). "Identifying regulatory networks by combinatorial analysis of promoter elements." Nat Genet **29**(2): 153-159.

Powell, J. R. and F. M. Ausubel (2008). "Models of Caenorhabditis elegans infection by bacterial and fungal pathogens." Methods Mol Biol **415**: 403-427.

Pujol, N., E. M. Link, et al. (2001). "A reverse genetic analysis of components of the Toll signaling pathway in Caenorhabditis elegans." Curr Biol **11**(11): 809-821.

Qin, Z. S., L. A. McCue, et al. (2003). "Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites." Nat Biotechnol **21**(4): 435-439.

Raharjo, I. and J. Gaudet (2007). "Gland-specific expression of C. elegans hlh-6 requires the combinatorial action of three distinct promoter elements." Dev Biol **302**(1): 295-308.

Rajewsky, N., N. D. Socci, et al. (2002). "The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons." Genome Res **12**(2): 298-308.

Romney, S. J., C. Thacker, et al. (2008). "An iron enhancer element in the FTN-1 gene directs iron-dependent expression in Caenorhabditis elegans intestine." J Biol Chem **283**(2): 716-725.

Saha, B., S. K. Singh, et al. (2006). "Activation of the Mitf promoter by lipid-stimulated activation of p38-stress signalling to CREB." Pigment Cell Res **19**(6): 595-605.

Sandelin, A. and W. W. Wasserman (2004). "Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics." J Mol Biol **338**(2): 207-215.

Schneider, T. D., G. D. Stormo, et al. (1986). "Information content of binding sites on nucleotide sequences." J Mol Biol **188**(3): 415-431.

Schones, D. E., P. Sumazin, et al. (2005). "Similarity of position frequency matrices for transcription factor binding sites." Bioinformatics **21**(3): 307-313.

Schulze, A. and J. Downward (2001). "Navigating gene expression using microarrays--a technology review." Nat Cell Biol **3**(8): E190-195.

Segal, M. R., K. D. Dahlquist, et al. (2003). "Regression approaches for microarray data analysis." J Comput Biol **10**(6): 961-980.

Shin, K. H., B. Choi, et al. (2008). "Analysis of C. elegans VIG-1 expression." Mol Cells **26**(6): 554-557.

Sifri, C. D., J. Begun, et al. (2005). "The worm has turned--microbial virulence modeled in Caenorhabditis elegans." Trends Microbiol **13**(3): 119-127.

Sifri, C. D., J. Begun, et al. (2003). "Caenorhabditis elegans as a model host for Staphylococcus aureus pathogenesis." Infect Immun **71**(4): 2208-2217.

Sinha, S., Y. Liang, et al. (2006). "Stubb: a program for discovery and analysis of cis-regulatory modules." Nucleic Acids Res **34**(Web Server issue): W555-559.

Sleumer, M. C., M. Bilenky, et al. (2009). "Caenorhabditis elegans cisRED: a catalogue of conserved genomic elements." Nucleic Acids Res **37**(4): 1323-1334.

Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol Biol Cell **9**(12): 3273-3297.

Spencer, W. C., G. Zeller, et al. (2011). "A spatial and temporal map of C. elegans gene expression." Genome Res **21**(2): 325-341.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

Stormo, G. D. and D. S. Fields (1998). "Specificity, free energy and information content in protein-DNA interactions." Trends Biochem Sci **23**(3): 109-113.

Stoyanov, C. N., M. Fleischmann, et al. (2003). "Expression of the C. elegans labial orthologue ceh-13 during male tail morphogenesis." Dev Biol **259**(1): 137-149.

Streit, A., R. Kohler, et al. (2002). "Conserved regulation of the Caenorhabditis elegans labial/Hox1 gene ceh-13." Dev Biol **242**(2): 96-108.

Sulston, J. E. and H. R. Horvitz (1977). "Post-embryonic cell lineages of the nematode, Caenorhabditis elegans." Dev Biol **56**(1): 110-156.

Sulston, J. E., E. Schierenberg, et al. (1983). "The embryonic cell lineage of the nematode Caenorhabditis elegans." Dev Biol **100**(1): 64-119.

Tan, M. W., S. Mahajan-Miklos, et al. (1999). "Killing of Caenorhabditis elegans by Pseudomonas aeruginosa used to model mammalian bacterial pathogenesis." Proc Natl Acad Sci U S A **96**(2): 715-720.

Tanay, A. and R. Shamir (2004). "Multilevel modeling and inference of transcription regulation." J Comput Biol **11**(2-3): 357-375.

Tavazoie, S., J. D. Hughes, et al. (1999). "Systematic determination of genetic network architecture." Nat Genet **22**(3): 281-285.

Taylor, J., S. Tyekucheva, et al. (2006). "ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements." Genome Res **16**(12): 1596-1604.

Teng, Y., L. Girard, et al. (2004). "Dissection of cis-regulatory elements in the C. elegans Hox gene egl-5 promoter." Dev Biol **276**(2): 476-492.

Timmons, L., D. L. Court, et al. (2001). "Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in Caenorhabditis elegans." Gene **263**(1-2): 103-112.

Tong, J. J., S. E. Schriner, et al. (2007). "Life extension through neurofibromin mitochondrial regulation and antioxidant therapy for neurofibromatosis-1 in Drosophila melanogaster." Nat Genet **39**(4): 476-485.

Trapnell, C., L. Pachter, et al. (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.

Trapnell, C., B. A. Williams, et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nat Biotechnol **28**(5): 511-515.

Troemel, E. R., S. W. Chu, et al. (2006). "p38 MAPK regulates expression of immune response genes and contributes to longevity in C. elegans." PLoS Genet **2**(11): e183.

Vazirani, V. V. (2001). Approximation Algorithms, Springer.

Wagmaister, J. A., G. R. Miley, et al. (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**(2): 550-565.

Wang, T. and G. D. Stormo (2003). "Combining phylogenetic data with co-regulated genes to identify regulatory motifs." Bioinformatics **19**(18): 2369-2380.

Wang, T. and G. D. Stormo (2005). "Identifying the conserved network of cis-regulatory sites of a eukaryotic genome." Proc Natl Acad Sci U S A **102**(48): 17400-17405.

Wasserman, W. W. and A. Sandelin (2004). "Applied bioinformatics for the identification of regulatory elements." Nat Rev Genet **5**(4): 276-287.

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

Whittle, C. M., E. Lazakovitch, et al. (2009). "DNA-binding specificity and in vivo targets of Caenorhabditis elegans nuclear factor I." Proc Natl Acad Sci U S A **106**(29): 12049-12054.

Zhang, Y., H. Lu, et al. (2005). "Pathogenic bacteria induce aversive olfactory learning in Caenorhabditis elegans." Nature **438**(7065): 179-184.

Zhao, G., N. Ihuegbu, et al. (2011). "Conserved Motifs and Prediction of Regulatory Modules in Caenorhabditis elegans." G3: Genes, Genomes, Genetics **Submitted**.

Zhao, G., L. A. Schriefer, et al. (2007). "Identification of muscle-specific regulatory modules in Caenorhabditis elegans." Genome Res **17**(3): 348-357.

Zhao, J., P. Wang, et al. (2007). "The C. elegans Twist target gene, arg-1, is regulated by distinct E box promoter elements." Mech Dev **124**(5): 377-389.

Zhao, Y. and G. D. Stormo (2011). "Quantitative analysis demonstrates most transcription factors require only simple models of specificity." Nat Biotechnol **29**(6): 480-483.

Zhao, Z., L. Fang, et al. (2005). "Distinct regulatory elements mediate similar expression patterns in the excretory cell of Caenorhabditis elegans." J Biol Chem **280**(46): 38787-38794.

Zugasti, O. and J. J. Ewbank (2009). "Neuroimmune regulation of antimicrobial peptide expression by a noncanonical TGF-beta signaling pathway in Caenorhabditis elegans epidermis." Nat Immunol **10**(3): 249-256.

CURRICULUM VITAE

Nnamdi Ihuegbu

nihuegbu@wustl.edu   **4309 Maryland Ave, Apt 10B
St. Louis, MO 63108**   **203-606-4699**

EDUCATION

Ph.D., Computational and Systems Biology   *Washington University School of Medicine,* St. Louis, MO

2005 – 2011 (October)

- Adviser: Gary Stormo, Ph.D.
- Thesis: Discovering conserved *cis*-Regulatory elements that regulate expression in *Caenorhabditis elegans*.
- Fellow, Kauffman Foundation for Life Science Entrepreneurship

M.Sc., Computer Science (Biology)   *Southern Connecticut State University*, New Haven, CT

2003-2005

- Adviser: Taraneh Seyed, Ph.D.
- Thesis: Computational Classification of Proteins Subcellular Localizations using Pattern Discovery Methods
- Fellow, Graduate Research Fellowship

B.Sc., Computer Science (Mathematics & Physics)   *Southern Connecticut State University*, New Haven, CT

1999-2003

- Advisers: Joseph Vitale, M.Sc. & Jason Stenzel, Ph.D.
- Senior Thesis: A Pipelined Approach to Solving the Generalized Born Surface Area Continuum Solvation Equations.
- Honors College Presidential Scholar

EXPERIENCE

Pre Doctoral Research Fellow – Laboratory of Gary Stormo, Ph.D.

Washington University in St. Louis, St. Louis, MO

September 2006 - October *2011*

- Developed and validated computational methods to efficiently and sensitively discover *cis*-Regulatory elements in complex eukaryotic genomes.

- Spearheaded collaboration with researchers at Harvard Medical School to discover a novel lead transcription factor involved in host response to *S.aureus* infection.

Teaching Assistant – Eukaryotic Genetics

January 2007 -May 2007

Washington University in St. Louis, St. Louis, MO

- Course Master: Douglas Chalker, Ph.D.

- Conducted laboratory workshops, assisted with lectures, grading, review sessions, and tutored students.

Research Cooperative – Pattern Discovery Group

*2004-200*7

IBM Research, Yorktown Heights, NY

- Designed a framework for evaluating and detecting novel microRNAs using their thermodynamic profiles

- Automated portions of existing approach to detect horizontally transferred genes in archaeal, bacterial, and viral genomes

- Designed interfaces to mine regulatory elements in whole-genome contexts as part of IBM Computational Biology toolkits called TEIRESIAS[®] and rna22[®].

- Integrated and optimized a prototype system for finding diagnostic patterns in patients' biomedical data as part of an IBM healthcare suite now called HealthMiner[®].

152

Data Analyst – Department of Education   May 2002 – August

Southern Connecticut State University, New Haven, CT   2005

- Analyzed data from surveys, focus groups, and other instruments and published in several federal government education-related evaluations.
- Assisted professors and students on research experimental designs and analysis.

Teaching Assistant – Applied Multiple

Regression/Correlation Analysis   May 2003-May 2004

Southern Connecticut State University, New Haven, CT

- Conducted laboratory workshops in SPSS for this graduate course.
- Assisted students with their research designs and tutored them on the implementation of appropriate statistical tests.

ADDITIONAL AWARDS & ACTIVITIES

- Winner, Olin School of Business IdeaBounce Elevator Pitch Competition (2011)
- Presenter, Rice Business Plan Competition (2011)
- Member, Olin Strategy and Consulting Association (2011) & Genetics Society of America (2009-present)
- Vice-President, BioEntrepreneurship Core (2009-2010)
- Mentor, Opportunities in Genomics Research & Young Scientist Programs (2008-2010)
- Graduate Student Graduate Assistantship Award (2003-2004)
- Outstanding Senior in Computer Science (2003) & Ida M. Cacesse Scholar (2000)
- President, People-To-People Club (2002)
- Fluent in English; Proficient in Igbo and French

PRESENTATIONS & PUBLICATIONS

- **Ihuegbu, N**, Foat, B, et. al (2009), *Modeling condition-specific gene expression using conserved cis*-regulatory elements. Presented at the 17th International C. elegans Meeting in Los Angeles and won 2nd Place Prize in Gene Expression Category out of 100 presenters.

- **Ihuegbu, N**, Visvikis, O, et. al (2011), HLH-30 is a novel transcription factor important for the C. elegans host response to S. aureus. Presented at the 18th International C. elegans Meeting in Los Angeles.

- **Ihuegbu, N**, Stormo, G, Buhler, J (2011), Fast, sensitive discovery of conserved genome-wide motifs. Submitted to the 8th RECOMB Regulatory Genomics Conference in Barcelona.

- **Ihuegbu, N**, Stormo, G, Buhler, J (2011), Fast, sensitive discovery of conserved genome-wide motifs. Accepted for publication to The Journal of Computational Biology.

- Zhang, G*, **Ihuegbu, N***, et al (2011), Conserved motifs and Prediction of Regulatory Modules in *Caenorhabditis elegans*. Submitted to G3: Genes, Genomes, Genetics.

- **Ihuegbu, N***, Visvikis, O*, et. al (2011), HLH-30/MiTF are novel transcription factors involved in host-defense response. Manuscript in preparation.

* Denotes equal contribution