Leveraging Genomic and Microarray Data to Find Direct Targets of the *C. elegans* Transcription Factor TBX-2

BY
Thomas James Ronan III
B.A., University of Chicago, 1991
J.D., Loyola University of Chicago, 2000
B.A., University of Missouri, 2003

THESIS
Submitted as partial fulfillment of the requirements
for the degree of Master of Science in Biological Sciences
in the Graduate College of the University of Illinois at Chicago, 2012

Chicago, Illinois

Defense Committee:

Dr. David Stone, Chair
Dr. Peter Okkema, Advisor
Dr. Theresa Orenic

This thesis is dedicated to my wife, Sarah, without whom I would not have had the opportunity to leave the practice of law for this new challenge, and to the other half of the T-box who has made the process of becoming a scientist so much fun.

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Dr. Peter Okkema, who gave me my first real opportunity to do scientific research, and who provided very much appreciated time and funding furthering my development as a scientist.

I would like to thank my thesis committee (past and present) — Dr. Teresa Orenic, Dr. David Stone, and Dr. Jennifer Schmidt — for their support in presenting my thesis and providing a classroom environment in which it was a pleasure to learn. I'd also like to thank Dr. Nava Segev for encouraging advice, key academic guidance, and wonderful support during my Ph.D. application process, Dr. Q. Tian Wang for always asking very difficult questions and keeping me on the right path, and Dr. Don Morrison for his excellent instruction and impeccable demeanor in and out of the classroom.

I would like to acknowledge the support of everyone in the Okkema lab, specifically Kalpana Ramakrishnan for sharing with me her wonderful and complex thoughts, Angenee Milton for great discussions about science and life, and Paul Huber for helping me 'Party Rock' my way through my time in the lab. I would also like thank Tanya Crum for her extensive time helping me with lab protocols, her thoughtfulness in discussions of research plans, and in helping me develop my understanding of efficient, well-directed, and worthy scientific research.

TR

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES (continued)

Figure **Page**

CHAPTER 1: BACKGROUND AND INTRODUCTION

1.1: Introduction

The process by which a single cell replicates and differentiates into a multicellular organism involves complex and coordinated temporal and spatial control of gene expression. As part of this process, networks of transcription factors influence the transcription of genes relevant to development. In these networks, key nodes control significant downstream events like developmental timing, cell-fate specificity, and organogenesis.

In one of these nodes lies TBX-2, a transcription factor key to the development of a major organ in *C. elegans*. In order to more clearly understand the function of TBX-2, we seek to identify direct downstream targets. A method of generating a subset of genes enriched for direct targets of TBX-2 is presented.

1.2: TBX-2 is a T-Box Transcription Factor

TBX-2 is a member of the T-Box family of proteins. T-Box proteins are a conserved family of DNA-binding transcription factors with a highly conserved DNA binding domain. T-box factors play key roles in animal development, and are responsible for proper heart and limb development in multiple organisms. Thus, studying TBX-2 in *C. elegans* has far-reaching implications for understanding the entire class of T-box factors in animal development.

1.3: T-Box Transcription Factors Are Involved In Developmental Regulation

T-Box transcription factors are involved in developmental regulation of a wide range of animals. (Showell et al., 2004) In mammals, T-box mutants exhibit defects in major internal organ systems, and limb development. (Stennard and Harvey, 2005; Hoogaars et al., 2007) In Drosophila, T-box mutants show defects in heart, limb, and wing development (Qian et al., 2008; Pflugfelder, 2009; Ryu et al., 2011). In each of these cases, the T-box protein acts as a transcription factor controlling a suite of downstream genes responsible for development.

The role of T-Box genes has most carefully been studied in heart development. In the mammalian heart, members of the Tbx1 and Tbx2 subfamilies play critical roles. Mutations in these proteins result in congenital heart defects resulting from improper differentiation, patterning defects, and improper organ development (Hoogaars et al., 2007). In the fly, *midline* and *neuromancer*, both homologs of TBX20, are critical for proper heart development and maintenance respectively. (Qian et al., 2008; Ryu et al., 2011)

1.4: In *C. elegans* TBX-2 is Critical for Organogenesis

In *C. elegans*, TBX-2 is required for proper development of the pharynx, the primary feeding organ (Roy Chowdhuri et al., 2006; Smith and Mango, 2007). *tbx-2(ok529)* animals lack the anterior pharyngeal muscle cells derived from the embryonic ABa blastomere (Sulston et al., 1983) ultimately resulting in larvae that arrest as they cannot feed due to malformed and unattached pharynxes.

tbx-2 is expressed strongly in the pharyngeal primordium during embryogenesis, as well as in seam cells, hypodermal cells, and in the tail (Roy Chowdhuri et al., 2006; Smith and

Mango, 2007)(A. Milton, P. Huber, and P. Okkema, personal communication). A partially penetrant temperature-sensitive allele, *tbx-2(bx59)*, results in predominantly wild type worms at permissive temperatures. At non-permissive temperatures, they present a similar phenotype to *tbx-2(ok529)* though with less severe defects and partial penetrance of the malformed and unattached pharynx phenotypes. *tbx-2(bx59)* animals resemble the weak phenotypes produced by *tbx-2(RNAi)* (Roy Chowdhuri et al., 2006). This interesting phenotype suggests that TBX-2 plays a key role in organ development.

The phenotype of TBX-2 mutants is partially characterized. While *tbx-2* mutants lack the anterior pharyngeal muscles cells, posterior pharyngeal muscles remain unaffected (Roy Chowdhuri et al., 2006; Smith and Mango, 2007). The anterior pharyngeal muscles are derived from a single early blastomere – ABa – while the posterior muscles descend from a separate lineage – MS (Sulston et al., 1983), suggesting that TBX-2 is required in the ABa lineage but is not critical in the MS lineage. Non-pharyngeal descendants of ABa develop normally, and some descendants of ABa express PHA-4, as all pharyngeal cells do (Smith and Mango, 2007). However, none of the ABa descendants expressing PHA-4 express the pharyngeal muscle marker CEH-22. Surprisingly, the missing cells from the lineage do not seem to switch to an alternate cell-fate or appear to undergo apoptosis (Smith and Mango, 2007).

1.5: Prediction of Direct Targets

Prediction of direct downstream targets of TBX-2 will help elucidate the role of *tbx-2* in organogenesis. Downstream targets can be predicted from expression data, and from DNA binding motifs. No sequence specific binding model currently exists for TBX-2 in *C. elegans* or

any other species. However, T-box proteins share a common, highly-conserved DNA binding domain. Sequence specific binding data exists for TBX5 and TBX20 in mice (Ghosh et al., 2001; Macindoe et al., 2009), and Brachyury in Xenopus (Kispert and Hermann, 1993). Crystal structures of T-box factors bound to DNA are available for TBX3 and TBX5 in humans (Coll et al., 2002; Stirnimann et al., 2010), and for Brachyury in Xenopus (Müller and Herrmann, 1997). These related proteins can provide a guide to prediction the binding behavior of TBX-2.

1.6: Overview of Process

Four independent methods were used to produce a data set enriched for direct targets of *C. elegans* TBX-2. In Chapter 2, a microarray experiment comparing wild-type and mutant *tbx-2* embryos was analyzed. The result of this analysis is a set of differentially-expressed genes which are enriched for genes downstream of *tbx-2*. In Chapter 3, a model for a class of T-box transcription factor binding sites was created in order to predict all likely binding sites in the *C. elegans* genome. Genes lacking binding sites in their regulatory regions were then ruled out as direct targets. In Chapter 4, an analysis of conservation of predicted T-Box binding sites was carried out in order to identify the most likely biologically functional sites. And in Chapter 5, an existing embryonic time course microarray data set was used to identify genes which demonstrate expression levels and timing compatible with being direct targets of TBX-2.

Each of these methods can be used independently to identify a subset of genes enriched for targets of TBX-2, but each method is limited in scope and each method has the capability to identify false targets. Both the methods in Chapter 2 and Chapter 5 rely on expression patterns in time or level, but lack information about the presence of binding sites through which TBX-2

could act directly. The method presented in Chapter 3 and improved upon in Chapter 4 relies on the presence of binding sites to predict direct targets, but is unable to consider expression information. Since the different methods used derive from different assumptions, the results of these methods should not share the same false positive predictions. Thus a combination of methods should provide a more reliable prediction. In Chapter 6, these methods are combined and should result in a high-quality pool of potential targets enriched for direct targets of TBX-2.

CHAPTER 2: TBX-2 WT-MUTANT MICROARRAY ANALYSIS

2.1: Introduction

The interesting phenotype for *tbx-2* suggests that it plays a key role in organ development. Discovering genes downstream of *tbx-2*, particularly direct targets, will help elucidate the mechanism by which it influences pharyngeal development in *C. elegans*.

A partially penetrant temperature-sensitive allele, *tbx-2(bx59)*, results in predominantly wild type worms at permissive temperatures, allowing for easy maintenance of homozygous animals. At non-permissive temperatures, they present a similar phenotype to *tbx-2(ok529)* though with less severe defects and partial penetrance of the malformed and unattached pharynx phenotypes. As *tbx-2(bx59)* animals also resemble the weak phenotypes produced by *tbx-2(RNAi)* (Roy Chowdhuri et al., 2006), *tbx-2(bx59)* is likely to be a hypomorphic loss-of-function allele.

Some T-box family members can act as either repressors or activators (Minguillon and Logan, 2003). While some members of the Tbx2 subfamily can act as activators (such as Tbx4 and Tbx5) other members can act as repressors (Tbx2 and Tbx3) (Habets et al., 2002; Lingbeek et al., 2002; Zaragoza et al., 2004). Since TBX-2 is the sole member of the Tbx2 subfamily in C. elegans, it is difficult to predict whether it will act as a repressor or activator, or to have dual-function. We know TBX-2 has repressive auto-regulatory activity as *tbx-2* promoter constructs are up-regulated in the mutant background (A. Milton, P. Okkema, personal communication).

TBX-2 also seems to act as a repressor for D2096.6 (L. Clary, P. Okkema, personal communication).

2.1.1: Summary of Analysis

In order to determine potential downstream targets of TBX-2, a microarray experiment was performed using Affymetrix *C. elegans* GeneChip arrays, comparing whole-genome expression profiles for wild type mixed-stage embryos and for *tbx-2(bx59)* mutant animals (L. Clary and P. Okkema, personal communication). The analysis of this experiment resulted in approximately 1400 differentially expressed genes (out of approximately 20,000 genes tested) at a false discovery rate of 5%. Of those differentially expressed genes, approximately 1179 genes increased expression levels in the *tbx-2(bx59)* mutants. This pool of 1179 genes represents a subset of all *C. elegans* genes enriched for potential direct targets of *tbx-2*, and are consistent with the hypothesis that TBX-2 is acting as a repressor.

2.2: Materials and Methods

2.2.1: Biological experiment design

A microarray experiment was performed comparing whole-genome expression profiles of two categories of mixed-stage embryos: wild type and *tbx-2(bx59)* mutant worms (L. Clary and P. Okkema, personal communication). Total RNA was isolated from N2 wild-type and *tbx-2(bx59)* mixed-stage embryo populations, and similar distributions of embryo stages were present in each population (L. Clary and P. Okkema, personal communication). Two N2 replicates, and three *tbx-2(bx-59)* replicates were prepared, and were sent to the Core

Genomics Facility (CGF) at the University of Illinois at Chicago for analysis using the Affymetrix *C. elegans* GeneChip microarray platform. Of these replicates, two wild type replicates (N2a and N2b) and two mutant *tbx-2* replicates (TBXb and TBXc) were ultimately used.

2.2.2: Affymetrix C. elegans GeneChip description

The Affymetrix C. elegans GeneChip was designed using the December 2000 release of the genome sequence, predicted transcripts, and EST sequences from the Sanger Center. The chip contains more than 22,500 probe sets which match 22,150 unique *C. elegans* transcripts (Affymetrix, 2002), and which map to approximately 20,000 unique WormBase Gene identifiers.

2.2.3: Analysis Parameters

The microarray data was analyzed using the R statistical programming language, using the Bioconductor suite of tools (Gentleman et al., 2004), and the Affy package. Normalization to correct for chip-to-chip variation was done using the Robust Multiarray Averaging (RMA) method of microarray normalization (Bolstad et al., 2003). Microarray results were pre-filtered using the *genefilter* function (25% of the probes have a measured intensity of at least 100 on the original scale and the coefficient of variation is between 0.7 and 10 on the original scale) (Chiaretti et al., 2004). The *limma* package (Smyth, 2004) was used to calculate differentially expression using the *limma* linear model fit, eBayes smoothing of standard errors, and Benjamini-Hochberg (BH) multiple test correction with a false discovery rate of 5% (Benjamini, 1995). Probes were matched to genes using the WormBase-provided Affymetrix-to-Wormbase-ID table (Harris et al., 2010) for WS210. Some probes mapped to more than one gene, and were

discarded. When one or more probes mapping to a gene were differentially expressed, that gene was considered to be differentially expressed.

Differentially expressed probes were visualized using heat maps created with the *heatmap2* Bioconductor package, and clustered using Euclidean distance measurements based on expression values. Gene Ontology term enrichment was calculated using the *GOStats* Bioconductor package using hypergeometric-based statistical tests ($p < 0.05$) both for the set of differentially expressed genes, as well as for 9 sub clusters generated via the *subclust* Bioconductor package.

2.3: Results

2.3.1: Quality assessment and removal of one replicate

Initial quality assessments were performed on the microarray data including inter- and intra-chip comparisons. The different GeneChips showed expected chip-to-chip variation (Figure 1), which was normalized using the Robust Multiarray Averaging (RMA) method of microarray normalization (Figure 2) (Bolstad et al., 2003). Expression levels between pairs of chips were compared, and the spiked-in control values were visualized. When comparing the two wild type replicates (N2a and N2b) or two mutant replicates (TBXb and TBXc) with each other, strong agreement between probes can be seen across the replicates, and spiked-in control values (red x marks) show very little change between replicates. Although variance is higher between N2 and *tbx-2* mutant replicates as expected, control values showed very little

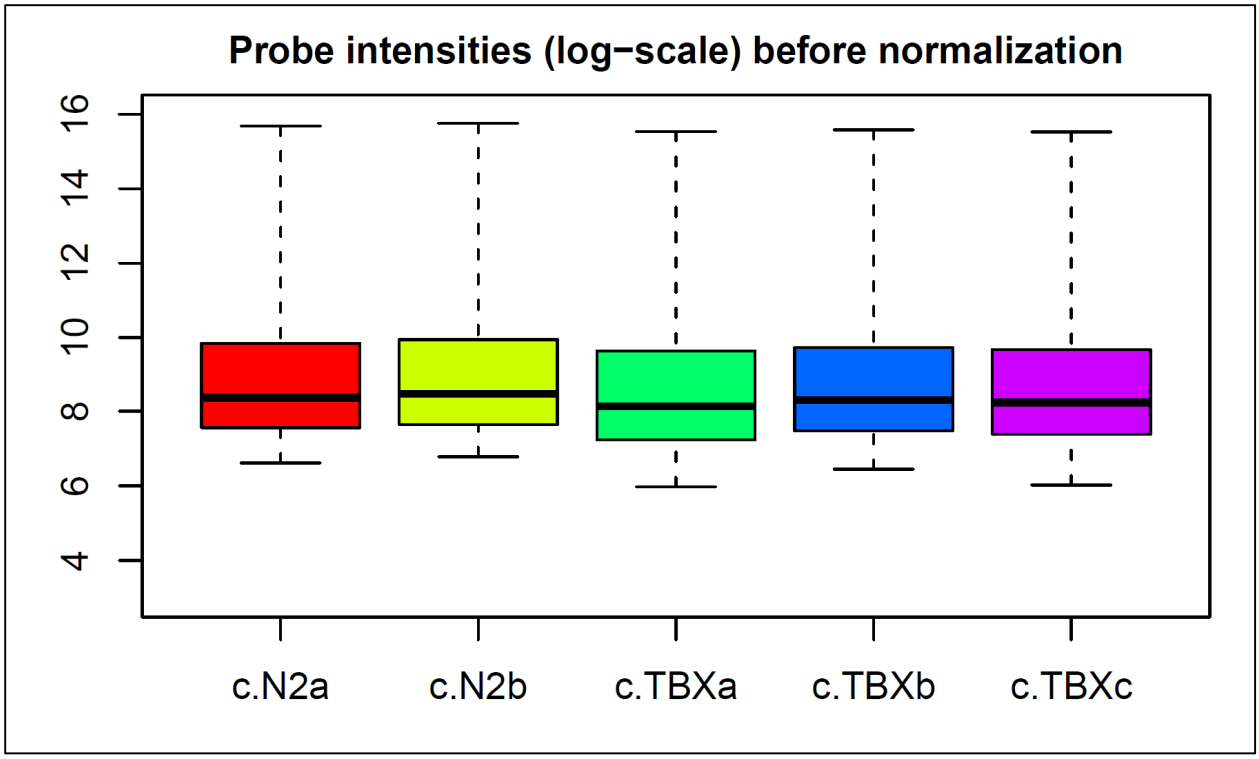

Figure 1: Boxplots of Intensity Values for the Five GeneChips Before Normalization
A boxplot of the probe intensity values for the five microarray GeneChips are shown before normalization. The black horizontal bar represents the mean, the colored box shows the limits of the first and third quartiles, and the whiskers outline the maximum and minimum values.

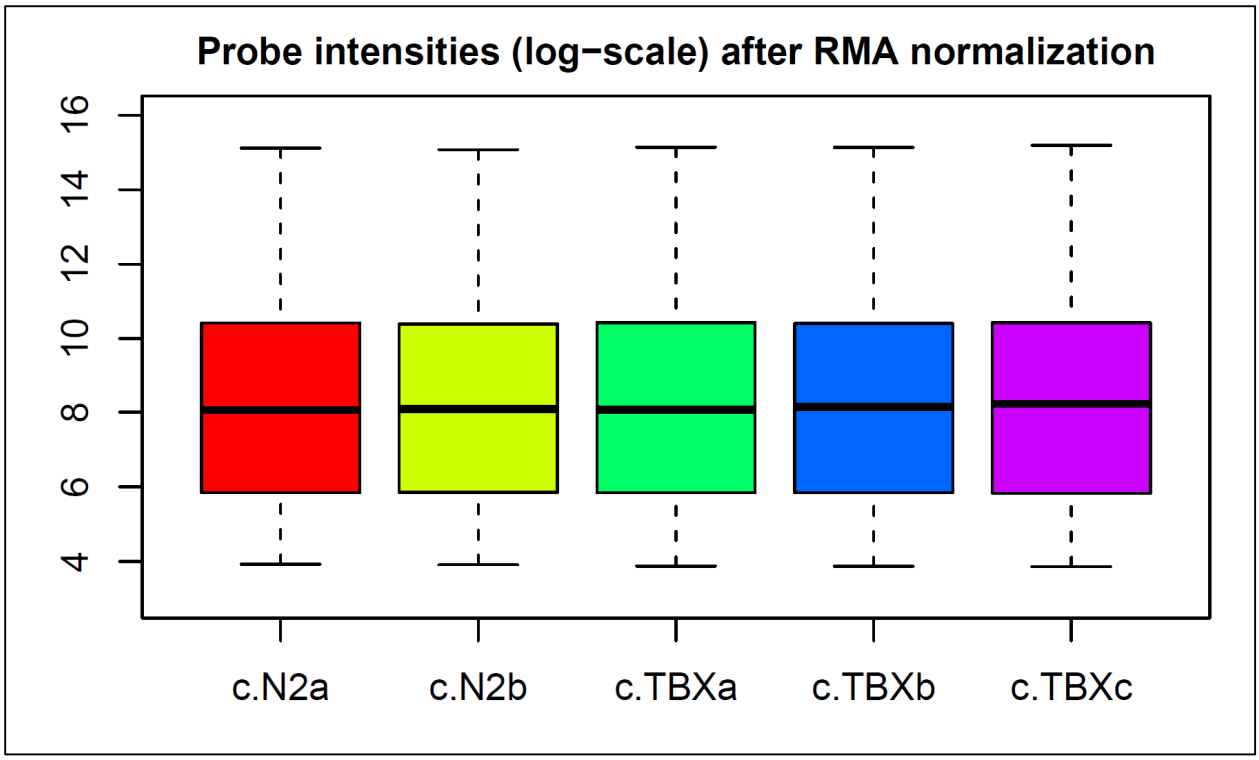

Figure 2: Boxplots of Intensity Values for the Five GeneChips After RMA normalization
A boxplot of the probe intensity values for the five microarray GeneChips are shown after RMA normalization. The black horizontal bar represents the mean, the colored box shows the limits of the first and third quartiles, and the whiskers outline the maximum and minimum values.

change when comparing the above four replicates with one another (Figure 3).

When comparing *tbx-2* mutant replicates with one another, one stands out as an outlier. The spiked in control values, which should measure the same on any chip, show more than 2-fold increase on the TBXa chip when compared with TBXb, TBXc, N2a, and N2b (Figure 4). This high variation indicates a significant problem with the TBXa replicate, and it was removed from further analysis.

2.3.2: Determination of Differentially Expressed Genes

The remaining four samples – two wild type samples (N2a and N2b) and two mutant *tbx-2* samples (TBXb and TBXc) – were analyzed to determine which genes were differentially expressed between the two categories. The limma package (Smyth, 2004) was used due to its improved performance with small numbers of biological replicates (Murie et al., 2009). Since a microarray analysis contains thousands of statistical comparisons, multiple test correction is required. Benjamini-Hochberg multi-test correction was used and the false discovery rate was set to 5% (Benjamini, 1995). Thus, at this high statistical threshold, no more than 5% of genes considered differentially expressed are expected to be false positives.

This methodology resulted in 1451 differentially expressed probe sets out of approximately 20,000 genes in *C. elegans* – with 1158 up-regulated in the mutant as compared to the wild type, and 293 down-regulated in the mutant as compared to the wild type (Figure 5). The Affymetrix oligo to Wormbase Gene map (WS210) (Harris et al., 2010) was used to map

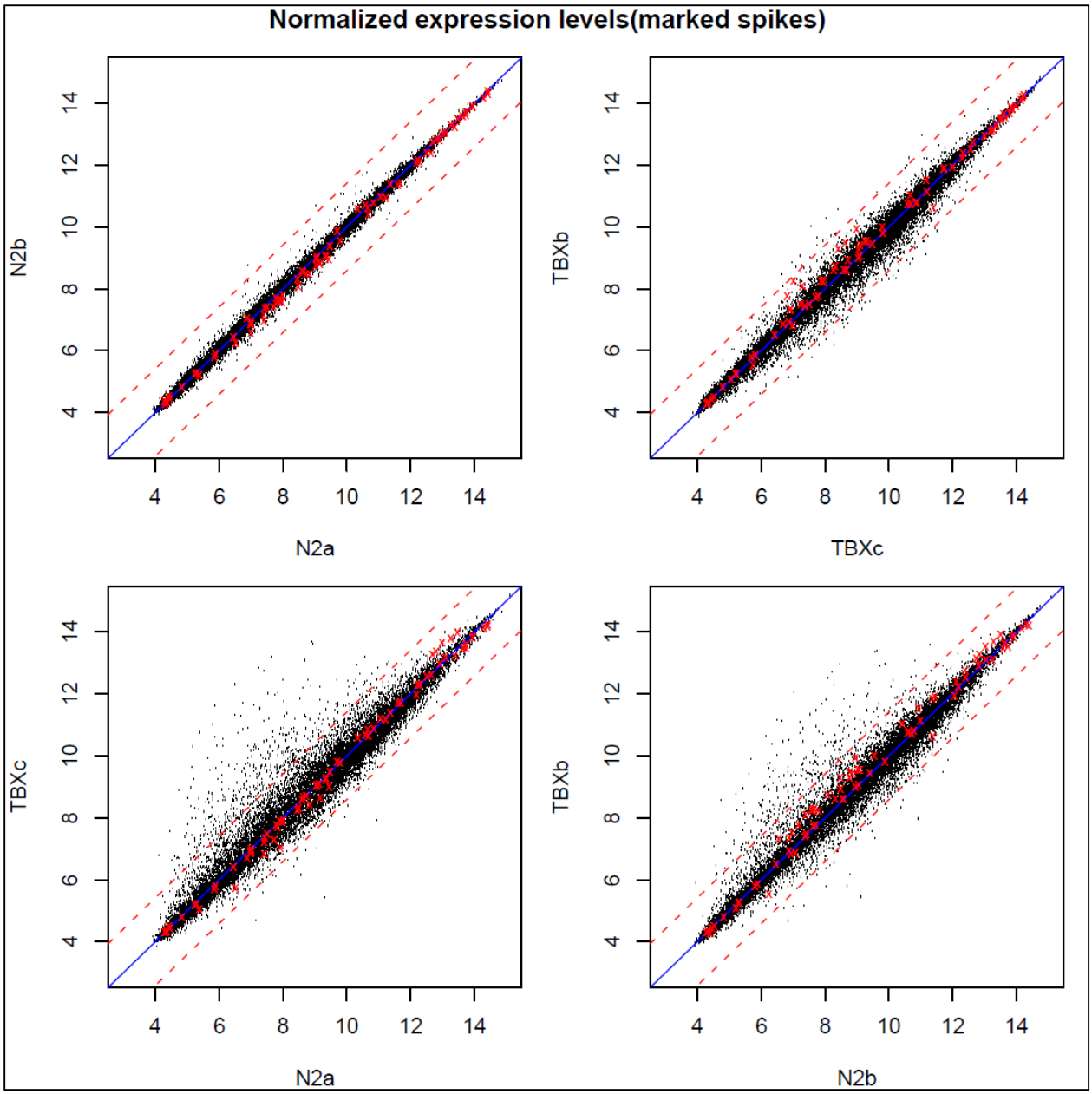

Figure 3: Pairwise Comparison of GeneChip Expression Levels for N2a, N2b, TBXb, and TBXc
Microarray replicate expression profiles are visualized using dotplots showing the distribution of normalized probe expression values by pairwise comparison between several microarray replicates. Scales are in normalized expression values. Black dots represent probe values, and the red x-marks represent spiked-in control values. The two wild type replicates N2a and N2b (upper left) show significant similarity and agreement between spiked-in values. The two mutant replicates TBXb and TBXc (upper right) show slight higher variation between replicates as evidenced by the wider distribution spread, but good agreement for spiked-in values. Comparisons between wild type and mutant replicates (lower left and lower right) show expected variation due to differential expression but strong agreement between spiked-in values.

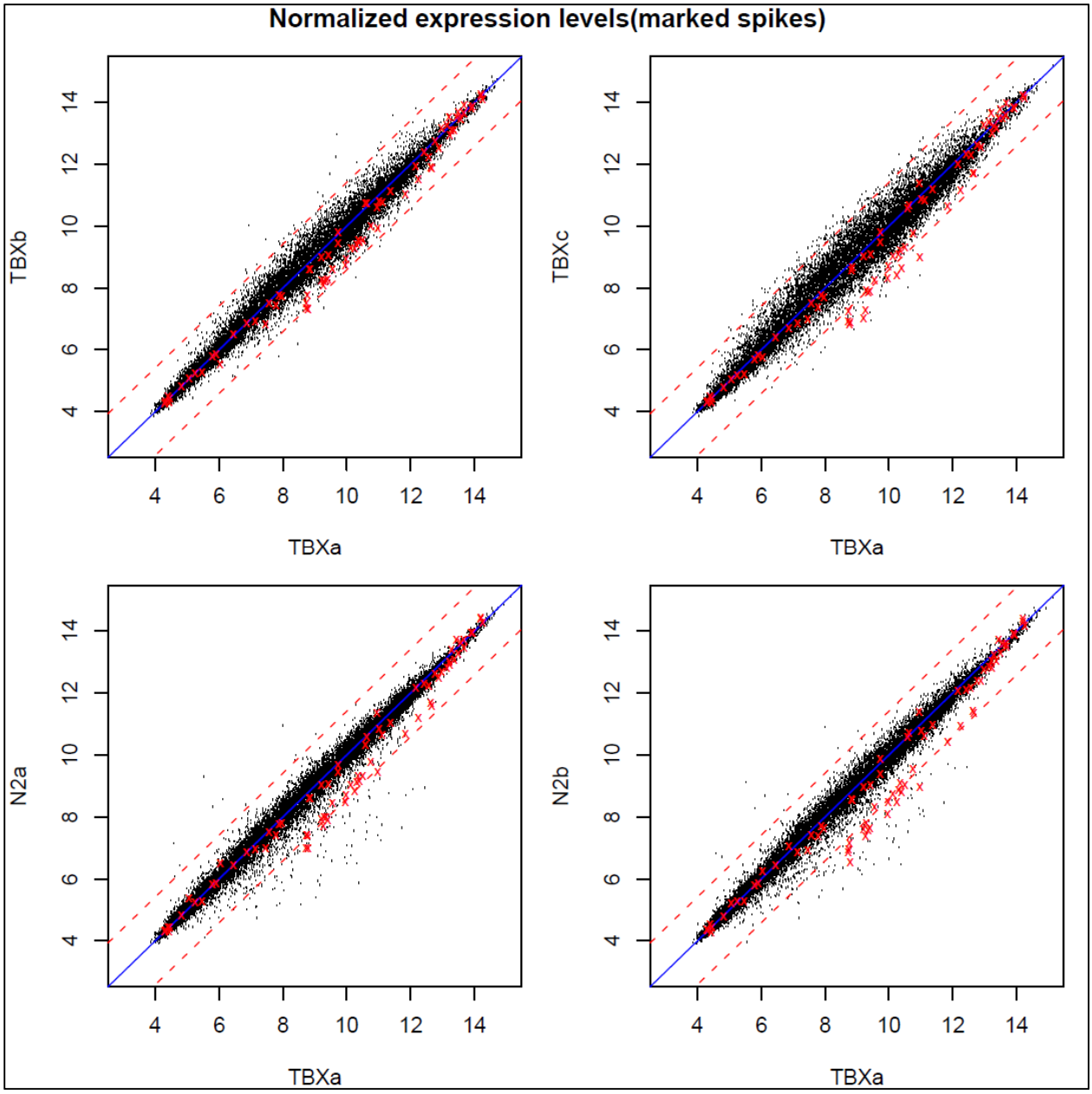

Figure 4: Pairwise Comparison of GeneChip Expression Levels for TBXa

Microarray replicate expression profiles are visualized using dotplots showing the distribution of normalized probe expression values by pairwise comparison between TBXa and other microarray samples. Scales are in normalized expression values. Black dots represent probe values, and the red x-marks represent spiked-in control values. TBXa (x-axis in all panels) shows high spiked-in values when compared to all other chips and low values of variance when compared to the wild type N2a (lower left) and N2b (lower right) replicates.

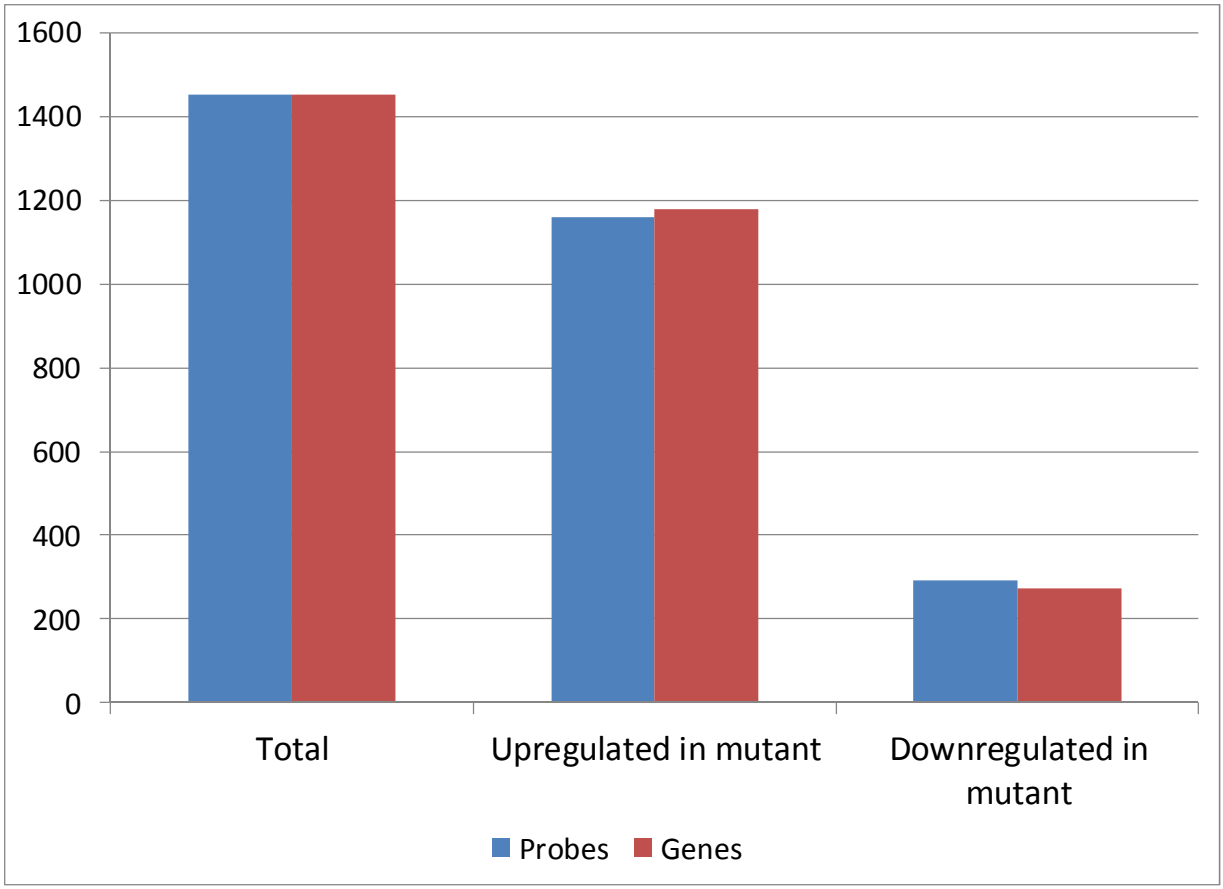

Figure 5: Differentially Expressed Probes and Genes
The number of differentially expressed genes and probes (at 5% FDR) are shown. Significantly more genes were up-regulated in the mutant than down-regulated in the mutant.

probes to genes. Probes and genes are not a one-to-one match. Some genes are mapped to by more than one probe. When one or more probes mapping to a gene were differentially expressed, that gene was considered to be differentially expressed. This resulted in 1452 differentially expressed genes, 1179 up regulated in the mutant as compared to the wild type, and 273 down regulated in the mutant as compared to the wild type (Figure 5).

2.3.3: Differentially Expressed Genes Are Visualized Using a Heat Map

In order to visualize changes in such a large number of genes, a heatmap was created using the *heatmap2* Bioconductor package. Probe expression levels are visualized using a color scheme to distinguish higher expression from lower expression, and probes are clustered on the map according to similar expression levels. Here, all 1451 probes are mapped and grouped by similar expression levels (Figure 6). Most of the differentially expressed probes change from lower expression levels (wild type on left) to higher expression levels (mutant on right) demonstrating significant up-regulation in the mutant. A small group of probes at the bottom show changes from higher expression levels to lower expression levels (from light green to dark green or from black to green on the graph) representing the small group of probes which are down regulated in the mutant (Figure 6).

2.3.4: Differentially Expressed Genes Are Characterized With Gene Ontology Terms

Gene Ontology (GO) Terms represent a prescribed vocabulary for describing biological annotations. Bioconductor was used to identify the set of GO-terms which are statistically enriched in differentially expressed genes.

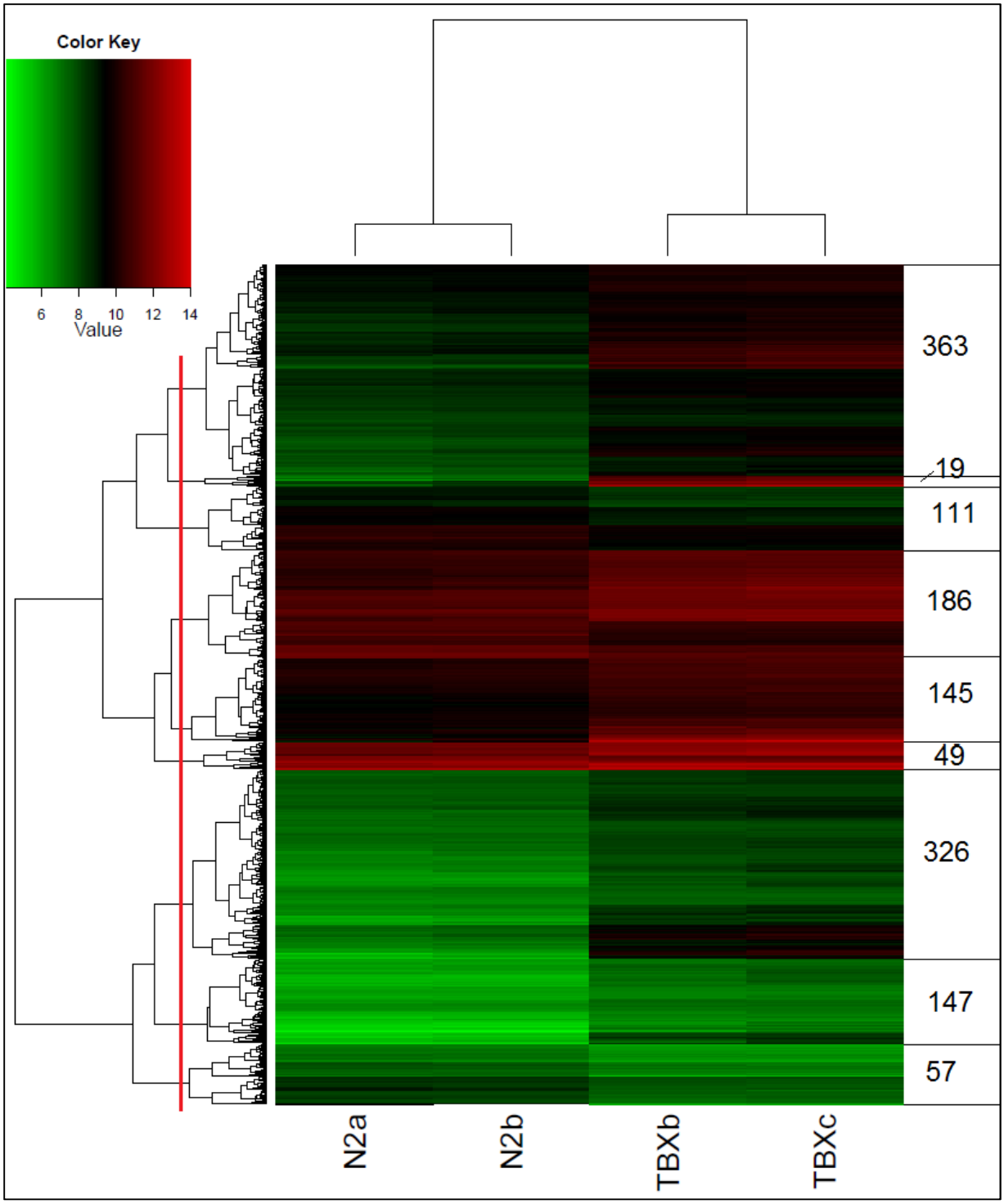

Figure 6: Heatmap of Differentially Expressed Genes

Normalized expression values for each differentially expressed probe (at 5% FDR) are visualized in a heatmap. Green represents lower expression levels, black represents middling values, and red represents high values (see key, upper left) for the normalized set. The horizontal red bar shows the cut location on the dendrogram from which the 9 clusters were created. Numbers (right) indicate the number of differentially expressed genes in each cluster.

In the cellular compartment category, several terms were enriched including the outer membrane-bounded periplasmic space, the periplasmic space, the cell envelope, the nuclear part, and the extracellular region. In the molecular function category, the most significant terms showed an increase in enzyme inhibitor activity, peptidase and endopeptidase inhibitor activity, as well as monooxygenase and oxidoreductase activity among other enzymatic activity. In addition, genes involved in cation binding, metal ion binding, and protein and receptor binding were enriched, along with genes affecting DNA polymerase activity (Appendix I). For enriched terms, the GO Term ID and go term are presented, along with the statistical significance (p-value) and the raw calculations on which enrichment is based: the actual count of the number of terms found for the gene subset (count) found, the expected number of terms considering the sample size (expected count), the number of terms in the category (size) and the log-odds-ratio calculation for those values.

On the whole, a wide range of biological processes seems to be accounted for in the differentially expressed gene set in mutant *tbx-2* animals.

2.3.5: Sub Clusters of Differentially Expressed Genes Demonstrate Differences in Statistically Significant Gene Ontology Terms

The differentially expressed genes were broken into 9 sub clusters, and GO-term enrichment was calculated for each sub cluster. These sub clusters showed significant differences in the character of enriched GO-terms. The 49-gene sub cluster, for example, showed enrichment of ATPase activity, chromatin binding, and nucleosome binding, as well as

enzymatic reactions involved with modifications of nucleic acid. The 326-gene sub cluster had enrichment genes for ion binding and membrane interaction (Appendix J).

Thus, clustering based on change of expression level results in some differentiation of the types of downstream effects.

2.4: Discussion

A significant number of genes from a wide-range of biological processes seem to be downstream of *tbx-2*. These genes are differentially expressed with high confidence based on the data examined and the high statistical thresholds set. Although a large number of downstream genes are affected in the mutant, not all of these are expected to be direct targets of TBX-2. The 1179 genes which are up regulated in the mutant background are candidates for further biological testing and bioinformatic analysis as potential direct targets of *tbx-2* acting as a repressor.

2.4.1: The Differentially Expressed Genes Represent Both Direct and Indirect Targets

This microarray experiment was designed to measure differential expression between two groups of organisms – wild type mixed-stage embryos, and *tbx-2(bx59)* mutant mixed-stage embryos. Thus, the result of the analysis is a list of genes whose expression changes above statistically significant thresholds, after correcting for multiple testing. These thresholds are high, resulting in a predicted 5% false discovery rate. In other words, one expects that no more than 5% of the genes predicted as differentially expressed will be changed due to random chance. The other 95% are expected to be due to differences between the wild type samples

and the mutant samples. Thus we have high confidence that these genes are differentially expressed in the mutant.

Nevertheless, the fact that these genes have a high confidence of differential expression does not mean that these are direct targets of TBX-2. While some are directly acted upon by TBX-2, due to the significant changes in a *tbx-2* mutant other differentially expressed genes are likely to be from downstream effects. Thus, the set of differentially expressed genes contains both direct targets and indirect targets.

2.4.2: The Differentially Expressed Genes Reflect Multiple Distinct Biological Processes, Implying Significant Downstream Effects

GO term analysis can be useful when attempting to understand differences in organisms due to mutations or disease. Enriched GO-terms highlight the changed processes in cellular systems when comparing a pool of healthy candidates to candidates with a disease, or when comparing a mutant to a wild-type organism.

In this case, a wide range of biological processes seems to be affected by genes downstream of TBX-2. Considering the significant physiological changes in the mutant, this is not unexpected. Although interesting, this information is not particularly helpful in identifying direct targets of TBX-2.

Sub clustering can be a most useful technique when analyzing microarrays with additional conditions, as one can elucidate subsets of genes that change in some, but not all, conditions and thereby understand effects in more detail. In this experiment, we have only two conditions, and clustering is done by expression levels. Nevertheless, some interesting patterns emerge. Downstream genes with similar expression levels seem to have enrichment for distinct

GO terms. Although it is not likely that these commonalities are directly due to similarities in level, genes with similar expression level changes may have upstream regulators in common. Thus, these groups of genes may represent clusters of downstream genes in the network of downstream effects. In each of these clusters, one may find one or more direct targets of TBX-2 influencing the expression levels of the entire group.

2.4.3: Down-regulated Genes

Since TBX-2 is known to act as a repressor, this microarray analysis (and all subsequent analysis using the microarray results) focuses on genes which are up-regulated in the mutant. However, T-box genes are known to be able to act as both activators and repressors, and some are known to have dual-function in different contexts. So, the set of differentially genes down-regulated in the mutant may represent activator function of TBX-2 rather than simply downstream indirect effects.

2.4.4: The Microarray Lacks Predictive Power When Used Alone, and Should Be Used In Combination With Other Methods

Analysis of a microarray experiment can provide a large amount of useful data, and provides downstream effects of a gene with high confidence. Nevertheless, most labs cannot test thousands of potential genes without great time and expense. In order to be useful for the purpose of finding direct targets of a particular gene, the results of a microarray should be combined with other analysis in order to improve the efficiency of biological research.

CHAPTER 3: PREDICTION OF TBX-2 TRANSCRIPTION FACTOR BINDING SITES

3.1: Introduction

3.1.1: TBX-2 is a Member of a Conserved Family of Transcription Factors

The protein TBX-2 is a T-box transcription factor. T-box proteins are a conserved family of DNA-binding transcription factors with a highly conserved DNA binding domain. Some members can act as either repressors or activators (Minguillon and Logan, 2003). T-box factors play key roles in animal development, and are responsible for proper heart and limb development in multiple organisms (Stennard and Harvey, 2005; Hoogaars et al., 2007). Thus, studying TBX-2 in *C. elegans* has far-reaching implications for understanding the entire class of T-box factors in animal development.

3.1.2: Transcription Factor Binding Site Prediction

A transcription factor is a DNA-binding protein that interacts with a sequence specific region of DNA, and which acts to influence transcription of nearby genes. This sequence specificity arises from the physical interaction of protein and nucleic acid, and is governed by the overall shape of the protein when folded as well as the position, charge, and shape of the residues which contact the DNA nucleotides and backbone. In essence, specificity is a physical property, but one which we can abstract into a characteristic of sequence.

In *C. elegans*, the regulatory regions of genes are often found within a short distance upstream of the transcription start site (Zhong et al., 2010; Niu et al., 2011). Transcription

factors bind to these regions, interact with other transcriptional machinery, and promote or suppress gene transcription. For transcription factors where the sequence specificity is known, potential binding sites are commonly predicted by scoring a particular DNA sequence against a model of the transcription factor binding specificity. These models vary in complexity from a simple consensus sequence – a text representation of the most commonly bound sequence – through models that incorporate probability into binding predictions. One of the most commonly used models is a position weight matrix (PWM) often represented by a sequence logo. The PWM is a matrix representing the independent probability of finding a particular nucleotide at a particular position, and the sequence logo is a graphical representation of that matrix. These models are used to predict potential binding sites by setting a threshold score and scanning the genome for matches to the model. A hit represents a prediction for physical interaction.

The predicted binding sites, along with their position, can be used identify direct targets of transcription factors. However, a large number of these predictions are known to be false positives (Wasserman and Sandelin, 2004). Although a transcription factor will tend to bind the sequence of its predicted binding site *in vitro*, it does not always do so *in vivo*. Other factors such as DNA accessibility from chromatin structure, binding site competitors, nearby binding proteins, and flanking sequences also influence whether or not these binding sites are biologically functional regulatory elements.

While the presence of a binding site can be used to predict a direct target, the absence of binding sites can also be useful for target prediction. While the presence of a binding site does not guarantee it is biologically relevant, the absence of any reasonably high-scoring

binding sites suggests that the gene is not a direct target of the transcription factor because there is no physically proximate site through which interaction is likely to occur to influence transcription.

3.1.3: Summary of Analysis

Here, a method is presented where the binding motif of a class of T-box transcription factors (including TBX-2) is predicted from structure and binding motifs of other, similar proteins. This motif is then used to scan the genome and map potential binding sites for a class of T-box transcription factors. Genes without sites in their regulatory regions are then eliminated as potential direct targets of TBX-2.

3.2: Materials and Methods

The TBXCore model was created by combining the position frequency matrices for the first eight bases of the Brachyury and TBX20 models with equal weights to generate a position frequency matrix (PFM). This matrix was then converted to a position weight matrix (PWM) (Wasserman and Sandelin, 2004). The composition of the *C. elegans* genome was calculated from the regulatory regions, defined as the intergenic region between the start codon of each gene, and the next exon or chromosome end upstream.

Protein sequences were aligned with ClustalW2 (Thompson et al., 2002) and visualized with Jalview (Waterhouse et al., 2009). Sequence logos were generated with the WebLogo sequence logo generator. (Crooks et al., 2004) The TBXCore model was used to search the genome using the JASPAR Perl API encoded in the TFBS module at a normalized score threshold

of 80% of the complete score range (Portales-Casamar et al., 2010) via a local copy of the *C. elegans* genome release WS220 (Harris et al., 2010) in a custom built MySQL database.

3.3: Results

3.3.1: Several T-box Factors Share Significant Similarities to TBX-2

TBX-2 in *C. elegans* shares significant sequence similarity to mouse TBX5 and TBX20, to Xenopus Brachyury, and human TBX3 (Figure 7), particularly in the DNA binding domain. While regions outside the DNA binding domain have diverged significantly, alignments of the DNA binding domain show very high conservation (Figure 8).

Crystal structure data for TBX3 and Brachyury demonstrate which residues of the T-box proteins contact either the DNA backbone or nucleotide bases when bound to a consensus sequence in DNA (Müller and Herrmann, 1997; Coll et al., 2002). These residues are highly conserved between TBX20 and Brachyury, with contacting residues completely conserved. Many surrounding residues are also completely conserved (Figure 8).

Brachyury, TBX20, and TBX3 have high sequence similarity in the DNA binding domain. TBX3 and Brachyury share similar structure in the DNA binding domain. As might be expected, the binding motifs are also very similar for TBX20 and for Brachyury. Both TBX20 and Brachyury bind to a similar core sequence with consensus AGGTGTGA, though they differ in 5’ flanking sequences (Kispert and Hermann, 1993; Macindoe et al., 2009). (Figure 9)

In contrast, TBX5 has similar sequence in the DNA binding domain but seems to have a very different DNA sequence specificity, with significant differences at positions 1 and 8 (Ghosh et al., 2001)(Figure 9). However, this difference in binding specificity can be partially explained

```
MTBX5/1-518      1 ------------MADTDEGFGLAR----------------TPLEPDSKDRSCDSKP----------  28
MTBX20/1-445     1 --MEFTASPKPQLSSRANAFSIAALMSSGGPKEKEAAENTIKPLEQFVEKSSCAQPLGELTSLDAHA  65
CelTBX2/1-423    1 --------------MAFNPFALGR----------------PDLLLPFMGAGVGGPG----------  26
XBRA/1-432       1 ----------------------------------------------MSATESCAKNVQYR----------  14
HTBX3/1-723      1 MSLSMRDPVIPGTSMAYHPFLPHRAPDFAMSAVLGHQPPFFPALTLPPNGAAALSLPGALAKPIMDQ  67

MTBX5/1-518     29 ----------ESALGAPSKSPSSPQAAFTQQGMEGIKVFLHERELWLKFHEVGTEMIITKAGRRMFP  85
MTBX20/1-445    66 EFGGGGGSPSSSSLCTEPLIPTTPIIPSEE--MAKIACSLETKELWDKFHELGTEMIITKSGRRMFP 130
CelTBX2/1-423   27 -AGGPPPNLFFSMLQAGFPPGPVGSPPEDDGVTDDPKVELDERELWQQFSQCGTEMVITKSGRRIFP  92
XBRA/1-432      15 ----------VDHLLSAVENELQAGSEKGDPTEKELKVSLEERDLWTRFKELTNEMIVTKNGRRMFP  71
HTBX3/1-723     68 LVGAAETGIPFSSLGPQAHLRPLKTMEPEEEVEDDPKVHLEAKELWDQFHKRGTEMVITKSGRRMFP 134

MTBX5/1-518     86 SYKVKVTGLNPKTKYILLMDIVPADDHRYKFADNK--WSVTGKAEPAMPGRLYVHPDSPATGAHWMR 150
MTBX20/1-445   131 TIRVSFSGVDPESKYIVLMDIVPVDNKRYRYAYHRSSWLVAGKADPPLPARLYVHPDSPFTGEQLLK 197
CelTBX2/1-423   93 AYRVKISGLDKKSQYFVMMDLVPADEHRYKFNNSR--WMIAGKADPEMPKTLYIHPDSPSTGEHWMS 157
XBRA/1-432      72 VLKVSMSGLDPNAMYTVLLDFVAADNHRWKYVNGE--WVPGGKPEPQAPSCVYIHPDSPNFGAHWMK 136
HTBX3/1-723    135 PFKVRCSGLDKKAKYILLMDIIAADDCRYKFHNSR--WMVAGKADPEMPKRMYIHPDSPATGEQWMS 199

MTBX5/1-518    151 QLVSFQKLKLTNNHLDPFGHIILNSMHKYQPRLHIVKAD--ENNGFGSKNTAFCTHVFPETAFIAVT 215
MTBX20/1-445   198 QMVSFEKVKLTNNELDQHGHIILNSMHKYQPRVHIIKKKDHTASLLNLKSEEFRTFIFPETVFTAVT 264
CelTBX2/1-423  158 KGANFHKLKLTNNISDKHGYTILNSMHKYQPRLHVVR----CADRHNLMYSTFRTFVFRETEFIAVT 220
XBRA/1-432     137 DPVSFSKVKLTN-KMNGGGQIMLNSLHKYEPRIHIVR--------VGGTQRMITSHSFPETQFIAVT 194
HTBX3/1-723    200 KVVTFHKLKLTNNISDKHGFTILNSMHKYQPRFHIVR----ANDILKLPYSTFRTYLFPETEFIAVT 262

MTBX5/1-518    216 SYQNHKITQLKIENNPFAKGFRGSDDLELHRMSRMQSKEYPVVPRSTVRHKVTSNHSPFSS------ 276
MTBX20/1-445   265 AYQNQLITKLKIDSNPFAKGFRDSS----------------------RLTDIERESVES------- 301
CelTBX2/1-423  221 AYQNEKVTELKIENNPFAKGFRDAGAGKREKKRQL-----------HRMNGDATQSPPG-------- 268
XBRA/1-432     195 AYQNEEITALKIKHNPFAKAFLDAKERNDYKDILDEGIDSQHSNFSQLGTWLIPNGGSLCSPN---- 257
HTBX3/1-723    263 AYQNDKITQLKIDNNPFAKGFRDTGNGRREKRKQLTLQSMRVFDERHKKENGTSDESSSEQAAFNCF 329

MTBX5/1-518    277 ----------ETRALSTSSNLGSQYQCENG-----------VSGPSQDLLPPPNPYPLAQEHSQI 320
MTBX20/1-445   302 ------------LIQKHSYARSPIRTYGEE---------------------------------- 319
CelTBX2/1-423  269 -----------KTASLPTHSPHPSESNSED------------------------------------ 287
XBRA/1-432     258 -----------PHTQFGAPLSLSSPHGCER------------------------------------ 276
HTBX3/1-723    330 AQASSPAASTVGTSNLKDLCPSEGESDAEAESKEEHGPEACDAAKISTTTSEEPCRDKGSPAVKAHL 396

MTBX5/1-518    321 YHCTKRKDEECSSTEHPYKKPYMETSPSEEDTFYRSG----------------------YPQQQGLS 365
MTBX20/1-445   320 -------DVLGEESQTTQSRGSAFTTSDNLS------------------------------LS 345
CelTBX2/1-423  288 -------DEPTLKKCKPEPSQTPTTSSLSTS------------------------------- 311
XBRA/1-432     277 ------YSSLRNHRSAPYPSPYTHRNNSPNN------------------------------LA 303
HTBX3/1-723    397 FAAERPRDSGRLDKASPDSRHSPATISSSTRGLGAEERRSPVREGTAPAKVEEARALPGKEAFAPLT 463

MTBX5/1-518    366 TSYRTESAQRQACMYAS------------------SAPPSEPVPSLEDISCN-------------- 399
MTBX20/1-445   346 SWVSSSSS-----------------------------FPGFQHPQP--------------------- 362
CelTBX2/1-423  312 -TTPTLSAHHP--------------------------LRSPQFCIP--------------------- 330
XBRA/1-432     304 DNSSACLSMLQS-----------------------HDNWSTLQMP--------------------- 325
HTBX3/1-723    464 VQTDAAAAHLAQGPLPGLGFAPGLAGQQFFNGHPLFLHPSQFAMGGAFSSMAAAGMGPLLATVSGAS 530

MTBX5/1-518    400 --------------------TWPSMPSYSSCTVTTVQPMDRLPYQHFSAHFTSG------------- 433
MTBX20/1-445   363 ---------------------------LTALGTSTASIATPIPHPIQGSLPPYS------------- 389
CelTBX2/1-423  331 ---------------------------PPIDMMYQNMPMDLLAHWQMATLFP---------------- 355
XBRA/1-432     326 -------------------------AHTGMLPMSHSTGTPPPSSQYPSLWSVS------------- 353
HTBX3/1-723    531 TGVSGLDSTAMASAAAAQGLSGASAATLPFHLQQHVLASQGLAMSPFGSLFPYPYTYMAAAAAASSA 597

MTBX5/1-518    434 ------------------------------------PLVPRLAGMANHGSPQLGEGMFQHQTSVAH 463
MTBX20/1-445   390 ------------------------------------RLGMPLTPSAIASSMQGSGPTFPSFHMPRY 419
CelTBX2/1-423  356 ------------------------------------QFSMALNSPAAAASLLS-KHLAKASSECKV 384
XBRA/1-432     354 ----------------------------------------NSAITPVSQSGGITNGISSQYLLGSTPH 381
HTBX3/1-723    598 AASSSVHRHPFLNLNTMRPRLRYSPYSIPVPVPDGSSLLTTALPSMAAAAGPLDGKVAALAASPASV 664

MTBX5/1-518    464 QPVVRQCG--PQTGLQSPGGLQPPEFLYTHGVPRTLSPHQYHSVHGVGMVPEWSENS--          518
MTBX20/1-445   420 HHYFQQGPYAAIQGLRHSSAVMTPFV-------------------------------          445
CelTBX2/1-423  385 EATSEDSEEAEKPEVKKEQKSVTP----PKKGGFDVLD--------LLSKP--------          423
XBRA/1-432     382 YSSLSHAVPSPSTGSPLYEHGAQTEIAENQYDVTAHSRLSSTWTPVAPPSV--------          432
HTBX3/1-723    665 AVDSGSELNSRSSTLSSSSMSLSPKLCAEKEAATSELQSIQRLVSGLEAKPDRSRSASP          723
```

Figure 7: Alignment of T-Box Protein Sequences

C. elegans TBX-2 (Q19691), mouse TBX5 (P70326) and TBX20 (Q9ES03), *Xenopus* Brachyury (P24781), and Human TBX3 (O15119) are aligned and colored using the ClustalW coloring scheme (Thompson, et al. 2002). The long region of high conservation (rows 2 through 5) is the DNA binding domain conserved in T-box proteins.

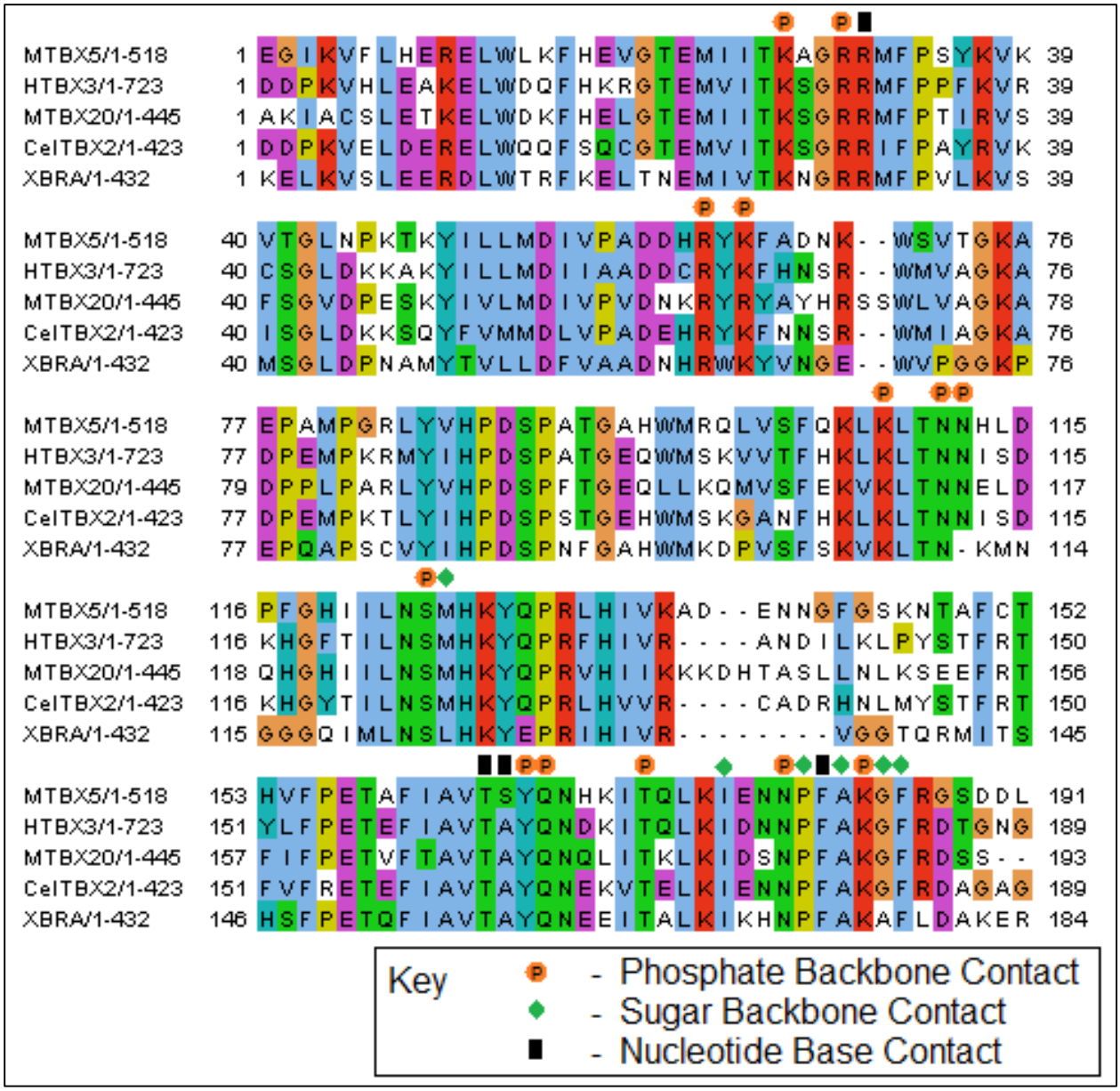

Figure 8: Annotated Alignment of Conserved Binding Domain in T-Box Protein Sequences
Here, a portion of the full alignment highlighting the DNA-binding domain is seen. Information from the structural data for Xenopus Brachyury and Human TBX3 is annotated. Each amino acid is labeled with the type of DNA contact, either phosphate backbone, sugar backbone, or a direct nucleotide base contact. All DNA-contacting residues show very high conservation, and all nucleotide base contacting residues show 100% conservation with one exception: an alanine to serine mutation in Mouse TBX5 at position 164 above (last row, second nucleotide base contact).

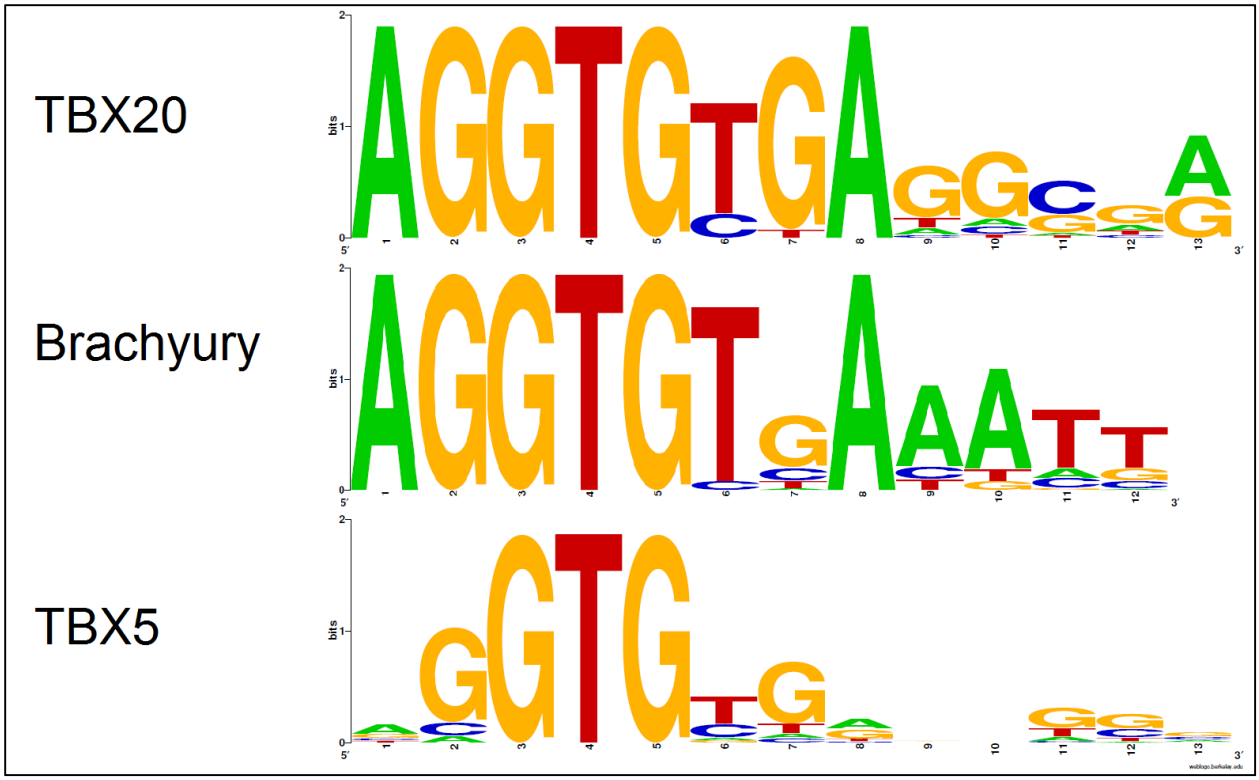

Figure 9: Sequence Specificity for T-Box Transcription Factors

Sequence logo representations of experimentally-generated binding motifs for TBX20, Brachyury, and TBX5 show the similarities in binding for positions 3 through 5 for all three proteins. TBX20 and Brachyury show a longer region of high similarity, from positions 1 through 8. The greatly diminished preference for adenine at position 8 correlates with the mutation in TBX5 for the amino acid likely to contact this nucleotide base.

by a key mutation at a residue which is likely to contact position 8. At position 165 in the mouse TBX5 DNA-binding domain, the amino acid contacting this nucleotide base has mutated from an alanine to a serine (Figure 8). The addition of the hydroxyl group could significantly affect binding as the charge and size are significantly different. The experimental sequence specificity for TBX5 shows a dramatic change at position 8 when compared to other T-Box proteins, but not at nearby positions 3, 4 or 5, and only a small change at position 7.

3.3.2: Building the TBXCore Model

No experimental structural or sequence specificity data exist for TBX-2. Nevertheless, due to the high sequence conservation, TBX-2 should adopt a similar structure in the DNA binding domain to that of Brachyury and TBX20. The protein sequence for TBX-2 is highly conserved in the DNA binding domain, and has perfect matches to TBX20 and Brachyury at a significant number of positions. Most importantly TBX-2 has 100% similarity at every position that contacts DNA in the TBX3 and Brachyury crystal structures (Figure 8). Furthermore, TBX-2 is more likely to be similar the TBX20 and Brachyury sequence specificity than to that of TBX5 because TBX-2 does not share the key alanine to serine mutation found in TBX5.

Although TBX20 and Brachyury share a common core sequence specificity in their first eight positions, the remaining several positions show significant differences. None of these remaining positions are known to contact DNA directly and may be influenced by other structural features not found in the crystal structures of the DNA binding domains. TBX-2 is likely to have a different binding specificity in these more distal positions from either TBX20 or Brachyury.

Thus, a model for a class of T-Box transcription factors most similar to TBX20, Brachyury, and TBX-2 can be represented by the first eight base pairs common to both binding motifs. This new model is called the TBXCore model (Figure 10). This model was created by combining the position frequency matrices which define the Brachyury and TBX20 models (data not shown) at the first eight positions with equal weights to generate a position weight matrix visualized by the sequence logo (Figure 10).

3.3.3: TBXCore Predicted Binding Sites

The TBXCore model was used to predict binding sites for this class of T-box proteins in the *C. elegans* genome in the regulatory region of each gene. The regulatory region was defined as the intergenic region between the start codon of each gene, and the next exon or chromosome end upstream. At a normalized threshold score of 0.80, approximately 103,000 TBXCore sites are predicted in the genome, and 15,000 genes have one or more TBXCore binding sites out of approximately 20,000 *C. elegans* genes. Conversely, approximately 5,000 genes do not have even one TBXCore binding site at this threshold, suggesting that these 5,000 genes (or approximately 25% of the genome) are not direct targets of any T-Box gene of this class including TBX-2.

3.4: Discussion

Attempting to predict complex interaction between a protein and DNA is a difficult task even with significant amounts of structural and biochemical data. Here a method is proposed to create a model that matches a broad class of T-box factors including TBX-2, but which attempts

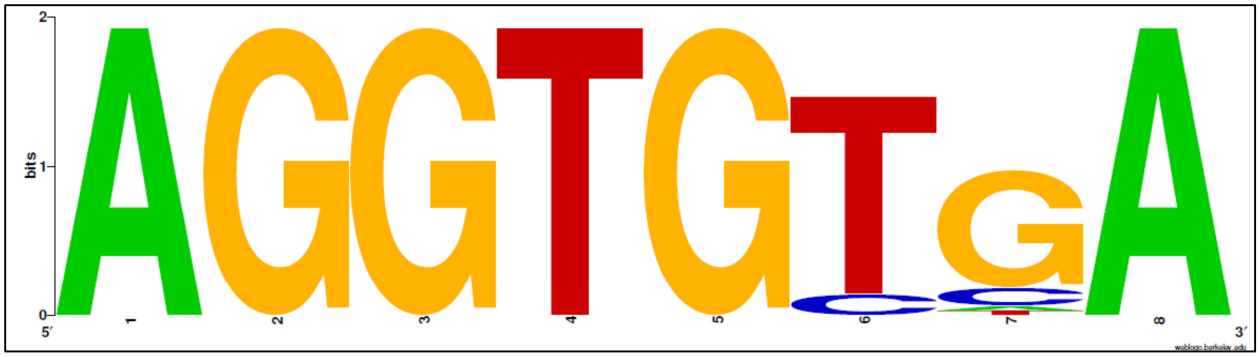

Figure 10: The TBXCore Sequence Logo
A sequence logo generated from the TBXCore position weight matrix.

to do so with minimal data specific to TBX-2, drawing predominantly from physical interactions and binding data of related proteins. This method is unlikely to be useful in the general case. However, in this specific case with a highly conserved DNA binding domain and significant overlap in sequence similarity and binding data, it seems reasonable to be able to draw limited conclusions from the data.

3.4.1: TBX-2 Should Bind The TBXCore Sequence

The first key conclusion to be drawn from the information available about T-Box transcription factor binding behavior is that TBX-2 is more likely to bind a sequence similar to TBX20 or Brachyury than a sequence like the sequence bound by TBX5. This conclusion relies on the fact that a key mutation at a nucleotide contacting residue seems to disrupt binding at that residue in TBX5, but this mutation is not present in TBX20, Brachyury or TBX2. All other DNA contacting residues are a perfect match, as are a large number of surrounding residues. The recent crystal structure of human TBX5 shows that the mutation from alanine to serine disrupts contact at position 8 of the binding motif (Stirnimann et al., 2010), supporting this conclusion. The most obvious difference in TBX5 is this key DNA contacting amino acid change. While this analysis fails to derive a binding specificity for TBX-2 directly, it rules out one of three known potential binding patterns.

The similarities between TBX20, Brachyury, and TBX3 are also a key factor in being able to predict binding behavior of TBX-2. They have very broad sequence identity and Brachyury and TBX3 have similar three dimensional structures in the DNA binding domain. All DNA contacting residues are a perfect match for Brachyury and TBX3, and the sequence specificity is

a perfect match for the first eight bases of TBX20 and Brachyury. Lacking empirical test data, since TBX-2 shows the same characteristics, it is reasonable to hypothesize that TBX-2 should bind the same sequence.

3.4.2: TBXCore Lacks Predictive Power When Used Alone, and Should Be Used In Combination With Other Methods

Using TBXCore eliminates 25% of the genome as unlikely to be direct targets of TBX-2. However, the remaining 75% of the genome may be targets of TBX-2, or of any similar T-Box transcription factor. Used alone, the TBXCore model lacks enough power to improve the efficiency of biological testing, or to directly predict direct targets of TBX-2. However, used in conjunction with a TFBS-independent method of predicting targets, it could be significantly more useful. For example, using this method in combination with a microarray experiment would provide complementary and synergistic results. While the microarray experiment provides direct experimental evidence of downstream targets, it fails to provide the link to direct targets. However, differentially expressed genes which lack TBXCore binding sites could be removed from consideration as direct targets and thereby improve the results of a microarray experiment with a simple bioinformatic analysis.

This method could be improved by increasing the likelihood that a particular predicted binding site is biologically relevant. Combining the results of the T0BXCore model with additional information about position and conservation of predicted binding sites should improve the biological relevance of a particular binding site, and thus should result in a more powerful predictor than the TBXCore model alone.

A better predictor of binding behavior would be experimental evidence of binding on a genomic scale, such as from the modENCODE project (Gerstein et al., 2010). Currently, genome wide chromatin immunoprecipitation data that is being collected for TBX-2 in *C. elegans* should provide high quality binding data with high biological relevance.

CHAPTER 4: CONSERVATION OF TBX-2 BINDING SITES IN 5' REGULATORY REGIONS

4.1: Introduction

4.1.1: Biological Significance of Predicted Transcription Factor Binding Sites

The TBXCore model has limited predictive power on its own, but the predictive power can be improved by adding an assessment of biological significance. Binding sites predicted from sequence motifs alone contain a significant number of 'false positive' predictions (Wasserman and Sandelin, 2004). While the transcription factor would be likely to bind the sequence *in vitro*, the predictions result in a high proportion of sites which are not biologically active (Brown, 2008), so these predictions are essentially technically correct, yet biologically irrelevant. When predicted binding sites are filtered for those with an improved likelihood of biological relevance, the usefulness of binding site prediction is greatly enhanced.

Biological relevance can be assessed using two factors: conservation and position relative to the transcription start site. Conservation provides information that a particular motif is unlikely to be present due to chance, thus improving the likelihood of biological relevance. The location of the binding site also provides information about its biological relevance in *C. elegans*. Recent experimental whole genome analyses of occupied transcription factor binding sites in *C. elegans* show regions of high occupation for a wide variety of transcription factor families within 500 base pairs upstream from the transcription start site of genes (Gerstein et al., 2010; Niu et al., 2011). Thus transcription factor binding sites in these regions should be more likely to be biologically relevant.

4.1.2: Conservation in Regulatory Regions

If a predicted binding site has survived evolutionary pressure and remains conserved in several related species, it is more likely to be a biologically active site rather than a random region of intergenic sequence (Brown, 2008). Conservation is typically defined as contiguous regions of sequence similarity in an alignment of related sequences.

In Figure 11, a conceptual representation of the genomic environment around the gene *mtl-2* can be seen as visualized by the UCSC genome browser. One conserved transcription factor binding site (site 1) and one non-conserved site (site 2) can be seen. Site 1 is found in a region of high conservation, as indicated graphically by the tall 'mesa' in the Conservation Track and also by the high sequence similarity in the 6-way Alignment track. Site 2, in contrast, has low conservation, and limited sequence similarity to the other species in the alignment. Site 1 would be considered 'classically' conserved, while site 2 would be considered non-conserved.

However, it is important to recognize that intergenic regulatory regions have different evolutionary constraints and pressures than coding regions. Coding regions are translated into proteins and selected for by physical function based on a three-dimensional structure. This structure is what is conserved, and since the structure is derived from the amino acid sequence, and the amino acid sequence is derived from the nucleotide sequence, then the nucleotide sequence is required to be preserved in order to preserve function. When a coding gene is present in widely diverging genomes, entire regions of similar nucleotide sequence are preserved in 5' to 3' order, since this order ultimately relates to the conserved physical structure and function.

Regulatory regions, in contrast, are not translated into proteins. Rather, short regions of

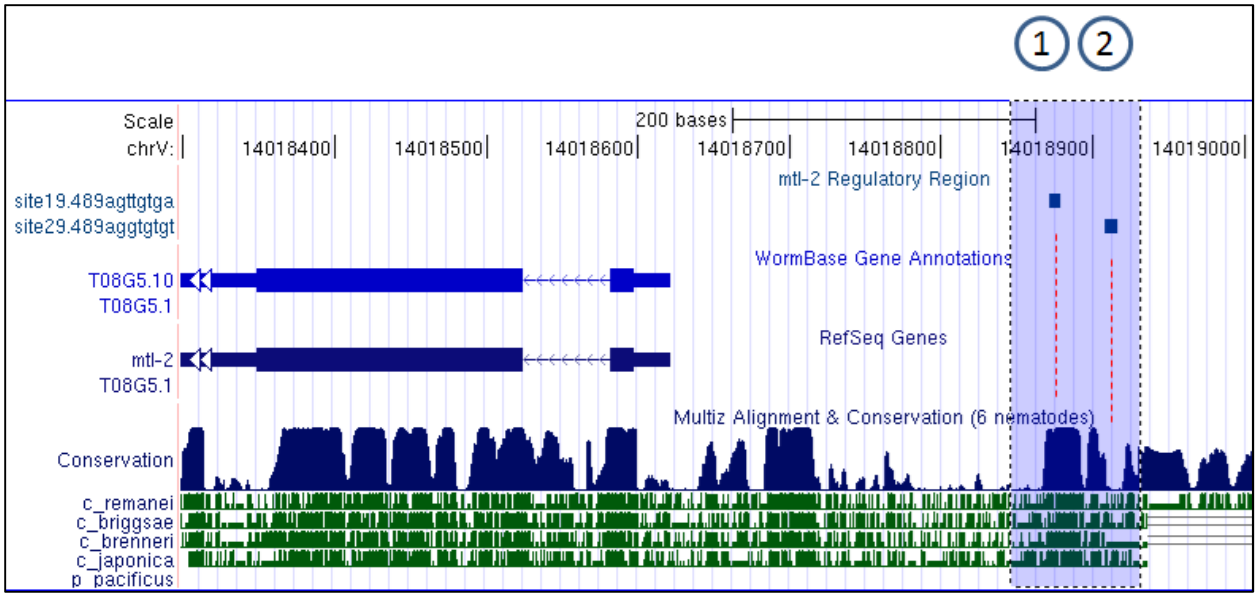

Figure 11: Binding Site Conservation Visualized by the UCSC Genome Browser

A region of *C. elegans* chromosome V containing the gene *mtl-2* is visualized with the UCSC Genome Browser. The *mtl-2* gene is on the reverse strand, and thus is shown right-to-left. Coding regions are indicated by dark blue horizontal bars and intronic regions by thin, hashed bars. Conservation is graphically indicated below, with the height of the graphic relating to phastCons conservation score (from 0 to 1). The blue shaded region on the right highlights an area in the 5'-intergenic region (the *mtl-2* regulatory region) containing two binding sites predicted with the TBXCore model. The sites are labeled 1 or 2 above the blue region. Site 1 shows high conservation, while site 2 shows low conservation.

the DNA are physically bound by proteins. Typically, multiple proteins interact with one another and with the transcriptional machinery to regulate transcription. The ability to bind and the ability to interact must be preserved. In contrast to protein coding regions, this interaction does not require a strict, ordered, lengthy region of sequence to be preserved. Rather, short regions of 5 to 20 base pairs must be conserved in linear order, but each of those short regions can be in different orders and can be in varying distances from one another while still preserving the required binding protein interactions. The sequence motif to which the protein is bound, and general proximity to other relevant motifs, is all that is required for a regulatory function to be conserved (Borok et al., 2009; Cameron and Davidson, 2009). Thus, the method of determining whether a regulatory sequence is conserved should account for the differing nature of regulatory regions from those of coding regions.

4.1.3: Summary of Analysis

Here, these criteria of regulatory conservation and proximity are applied to the predicted transcription factor binding sites from the TBXCore model and are used to refine the prediction to increase the biological relevance of each predicted site.

4.2: Materials and Methods

Regulatory regions for each gene were defined as the intergenic region between the start codon of a gene, and the next exon or chromosome end upstream. These regulatory regions were calculated for five complete *Caenorhabditis* genomes (WS210): *C. elegans*, *C. briggsae*, *C. brenneri*, *C. remanei*, and *C. japonica*. These regions were then searched with the TBXCore model using the JASPAR Perl API encoded in the TFBS module at a normalized score

threshold of 80% of the complete score range (Portales-Casamar et al., 2010) via a local copy of the *Caenorhabditis* genomes in a custom built MySQL database.

For each predicted binding site in *C. elegans*, conservation was determined in two ways. Classically defined conservation was measured by a quantitative phastCons conservation score extracted from the UCSC Genome Browser and a qualitative assessment of regulatory conservation called 'input conservation' was derived from the presence of binding sites in regulatory regions from homologous genes.

PhastCons scores were derived using the May 2008 (WS190/ce6) assembly in the UCSC Genome Browser, and the phastCons (phastCons6way) track was extracted using the genomic coordinates in the *C. elegans* genome for each binding site. Scores for the site were then averaged to determine the conservation score. Sites with scores above 0.80 were considered conserved. Thus, if a gene in C. elegans has one or more binding site with phastCons scores greater than 0.80, that gene was considered have a 'classically' conserved binding site in its regulatory region.

An additional measure of conservation, 'input-conservation', was determined for each gene. For each *C. elegans* gene with one or more binding sites, the binding sites in homologous genes in other nematode species were assessed. If one or more homologous genes a *Caenorhabditis* species had a binding site, the gene was considered to have a conserved binding site in that species. The number of species with conserved binding sites was then tallied and this is the input-conservation score. Thus, the input-conservation score can range from a 0 (representing a gene with TBXCore binding sites only in *C. elegans*) to a 4 (representing a gene

with binding sites present in four nematode species). (See Figure 12 for an example of the gene *lea-1* with an input-conservation score of 3.)

The distance of each binding site to the start codon in the regulatory region was determined. The distance was calculated from the 5'-most position of the binding site, to the 5'-most position of the start codon independent of strand. For each gene in *C. elegans* with one or more binding sites, the distance of the most proximate binding site was calculated. For *C. elegans* genes with the closest binding site within 1000 base pairs of the start codon, the closest binding sites in homologous genes for other nematode species were assessed. If one or more homologous genes in a *Caenorhabditis* species had a closest binding site within 1000 base pairs of the start codon, the gene was considered to have a proximate input-conserved binding site in that species. The number of species with proximate input-conserved binding sites was then tallied and this is the proximate input-conservation' score. Thus, the proximate input-conservation score can also range from a 0 to a 4. (See Figure 12 for an example of the gene *lea-1* with a proximate input-conservation score of 3.)

4.3: Results

4.3.1: Classically conserved TBXCore binding sites

The TBXCore model alone predicted approximately 103,000 binding sites for 15,000 genes at a normalized threshold score of 0.80. Of these, only 13,428 binding sites have a phastCons conservation score above 0.8, representing approximately 5,000 genes with

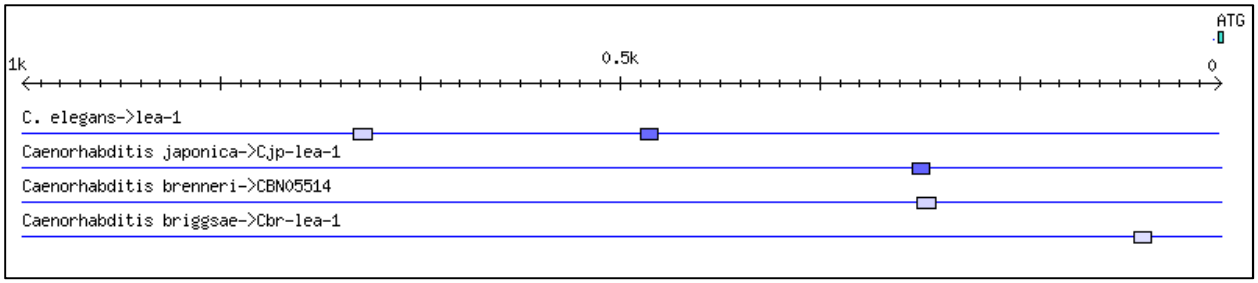

Figure 12: Graphical Representation of Input Conservation

The regions 1000bp upstream of the start codon were compared for the lea-1 gene in *C. elegans* and orthologs. Predicted TBXCore binding sites were mapped (blue boxes) onto the regions with the start codons for each gene aligned at the right. The quality of the binding site (above the minimum 80% threshold score) is indicated by the color of the box. Light blue boxes represent scores closer to 80%, while darker blue boxes represent sites with scores closer to 100%. Species and orthologs without binding sites are not shown. Here, lea-1 shows two binding sites in *C. elegans*, one of which is a very high scoring site. The gene *Cjp-lea-1* in *C. japonica* shows one high scoring binding site. Orthologs in *C. brenneri* (*CBN05514*) and *C. briggsae* (*Cbr-lea-1*) show one binding site each. The input conservation score (representing the number of species with orthologs having binding sites present in the regulatory region) for this gene is 3. The proximate input conservation score (representing the number of species with orthologs having binding sites present within 1000bp of the start codon in the regulatory region) for this gene is also 3.

classically conserved binding sites in the *C. elegans* genome. Thus, requiring the presence of one or more conserved TBXCore binding sites eliminates approximately 75% of *C. elegans* genes as likely direct targets of TBX-2 (Figure 13).

Adding the requirement that conserved binding sites be proximate to the transcription start site further reduces the potential direct target pool. Only 3,659 binding sites are within 1000bp of the start codon, and have a phastCons conservation score above 0.8, representing approximately 2800 genes with proximate classically conserved binding sites in the *C. elegans* genome (Figure 13). This additional requirement eliminates 86% of *C. elegans* genes as unlikely direct targets of TBX-2.

4.3.2: Description of 'Input Conserved' TBXCore binding sites

Input-conservation scores were calculated for all *C. elegans* genes. Of approximately 20,000 genes, 6855 genes show the highest input-conservation score of 4. These genes have a TBXCore binding site present in the regulatory region of one or more orthologs in the 4 *Caenorhabditis* species tested. Similarly, 12,528 genes have a score of 3 or more, 16,209 genes have a score of 2 or more, and 18,351 have a score of 1 or more (Figure 14).

4.4: Discussion

4.4.1: Classical conservation under-predicts biologically active TFBS

Alignment-based methods of determining conservation, such as presented in the multiz6way alignment visualized by the UCSC Genome Browser, are most relevant for coding

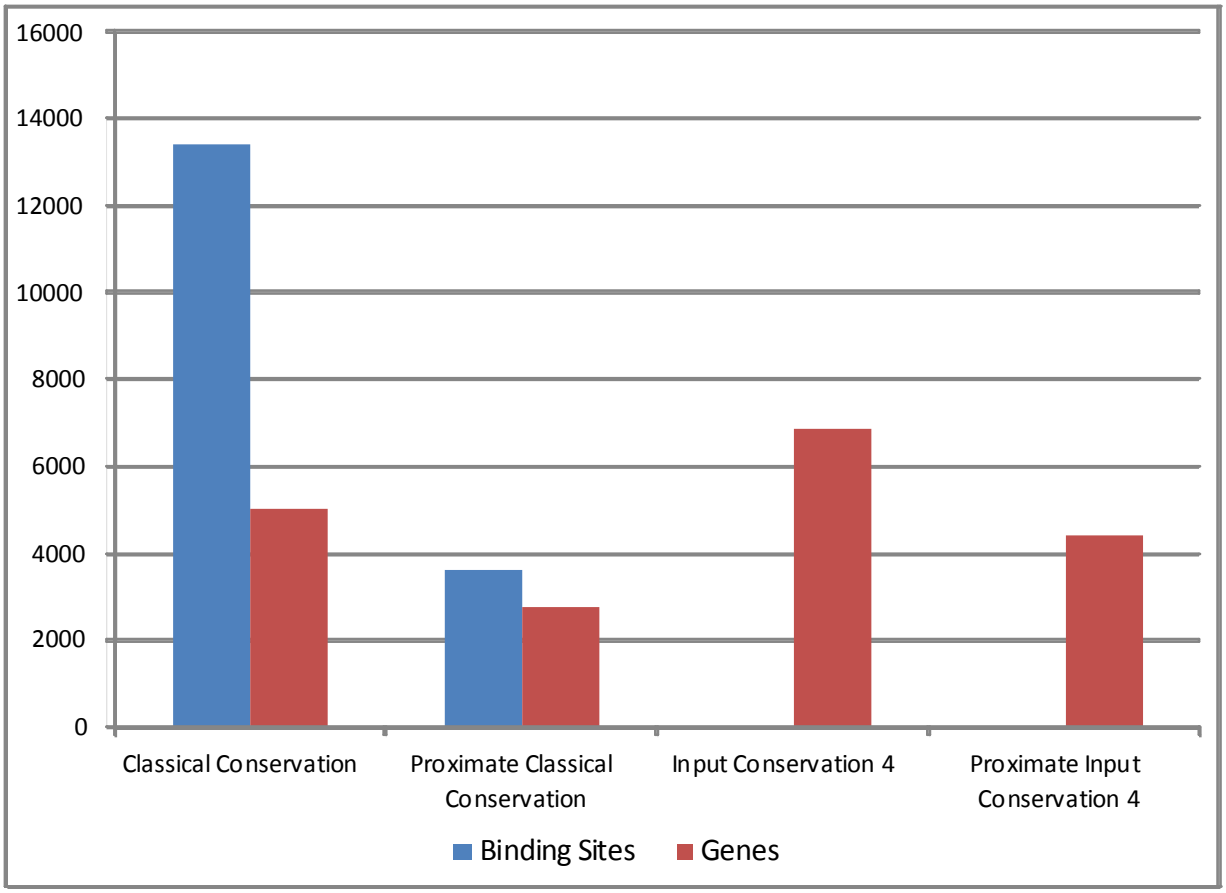

Figure 13: TBXCore Distribution by Conservation Type

The number of conserved binding sites predicted by the classical conservation method and the input conservation method are shown. Classical conservation is based on binding sites having phastCons scores above 0.80. An input conservation score of 4 requires all four nematode species to have one or more orthologous genes with binding sites in the regulatory region. Proximate conservation restricts the analysis to binding sites within 1000bp of the start codon. (As input conservation is a gene-level measurement, numbers are not available for binding sites.)

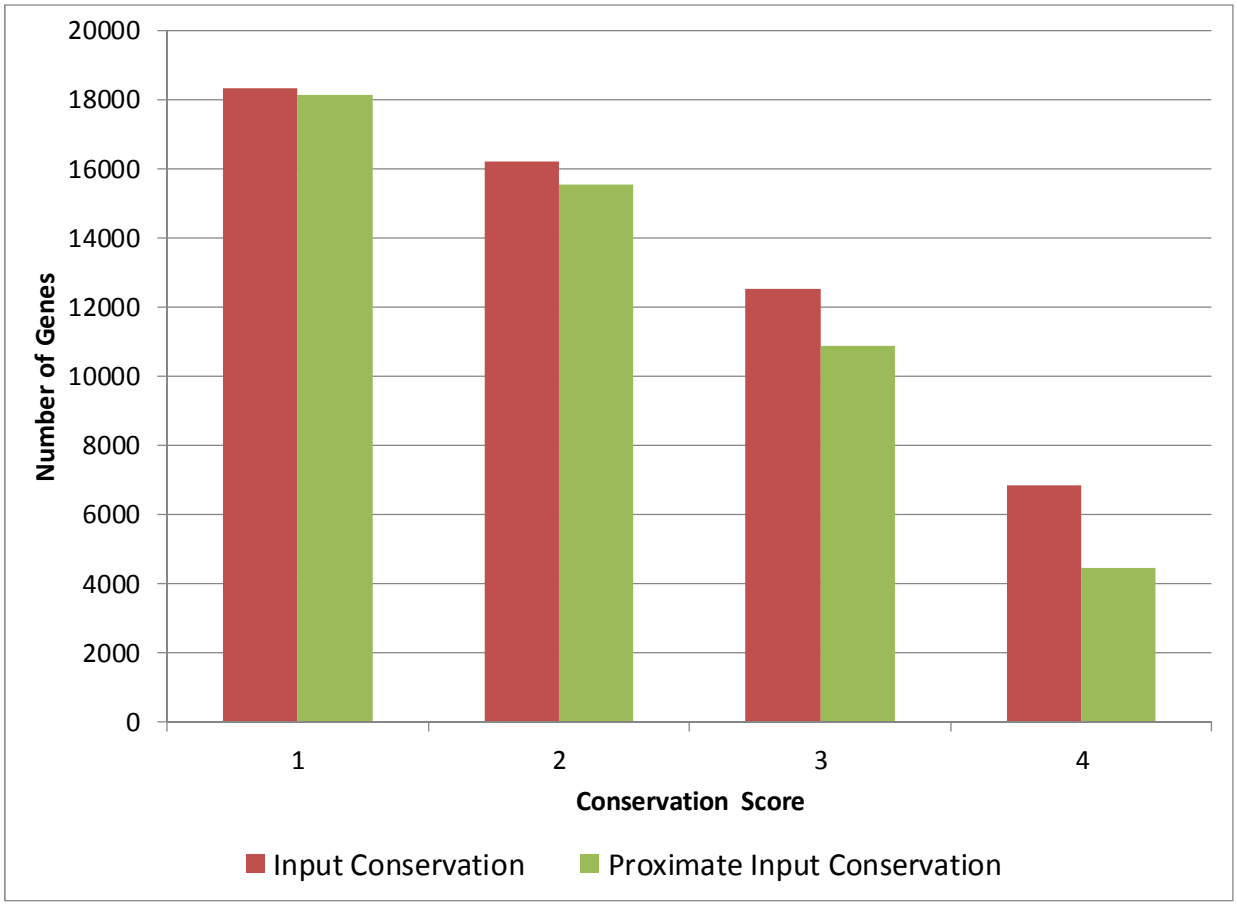

Figure 14: Distribution of Input Conservation Scores

Numbers of genes for input conserved binding sites are compared for various input conservation scores, and for proximate conservation (counting only genes with one or more binding sites within 1000bp of the start codon). At higher conservation scores, the ratio of proximately input conserved genes to input conserved genes decreases.

regions of the genome. While classically conserved regulatory regions also demonstrate conservation of function, the extent of functional conservation is under predicted. Extensive documentation exists in multiple organisms for biologically relevant binding sites which are present in orthologous regulatory regions, but which are not classically conserved. (Cameron and Davidson, 2009) None of these sites can be found with a method relying solely on alignment to determine conservation.

4.4.2: The Input Conservation Method Yields More Binding Sites Than Classical Conservation While Maintaining Criteria for Biological Relevance

The choice of the name 'input-conservation' comes from the idea that biologically relevant binding sites act as inputs to the control mechanism of a gene, and it is the presence of binding sites in the region in multiple species that indicates the conservation of the input, rather than classical conservation from an alignment.

While all classically conserved sites are input-conserved, not all input-conserved sites are classically conserved. Using classical conservation alone will find conserved binding sites and thus improve binding site predictions for biological relevance, but the method is under-inclusive. In *C. elegans*, implementation of the input-conservation method at a score of 4 resulted in an increase from 5069 classically conserved genes to 6855 input-conserved genes, an increase of approximately 35% over the classical conservation method. Implementation of the proximate input-conserved method resulted in an increase from 2805 classically conserved genes to 4460 input-conserved genes, an increase of approximately 59% over the classical conservation method. While one goal of these methods is to reduce the number of candidates for testing, a competing goal is to make the most accurate prediction of the number of

potential direct targets. Restricting the analysis to only classically conserved targets causes a significant number of potential direct target genes to be excluded.

Thus, further analysis is conducted with input-conservation as the primary method of conservation analysis.

4.4.3: Conservation Analysis Improves The Predictive Power of the TBXCore Model, and Can Be Used In Combination With Other Methods

Conservation analysis enriches the pool of potential direct targets for biologically relevant binding sites. While it is likely that some biologically active binding sites are excluded using any conservation method, the likelihood that the remaining pool will contain relevant sites is greatly enhanced. Used in combination with the results of an independent analysis, such as a microarray experiment, could greatly improve the power of either analysis alone.

CHAPTER 5: EMBRYONIC GENE EXPRESSION ANALYSIS USING TIME-LAGGED ANTI-CORRELATION (TILAC)

5.1: Introduction

5.1.1: Predicting Targets With Time Course Microarray Data

Differential gene expression, as characterized by a microarray experiment comparing wild type and mutant animals, is useful to determine genetic changes downstream of the mutant genes. The changes measured occur *after* the effects of the mutant gene are realized. Thus, temporal changes are implicit in the concept of this type of microarray experiment, though in a limited fashion. Resolution is limited to *before* and *after*, and no data is gathered on which organisms, which cells, or during which specific time these changes occur.

Time-course microarray experiments measure expression levels at multiple time points. So while still lacking organism-level or cellular-level resolution, the resolution is improved over the two-condition microarray experiments, and differences in expression at each point in time are resolvable. Thus, one has access to information which can help determine not only that a target is downstream, but in combination with rates of transcription and translation, one can use the time-scale to determine how far downstream a target is. A target changing in the correct way and at the correct time is a candidate for a direct target.

A typical measure of relatedness relies on a distance calculation generated from correlation between two time-series data sets using the Pearson correlation coefficient (PCC). Unlike measures of Euclidean distance, the PCC weights direction rather than absolute value of

change, and is useful for gene expression analysis. High PCC scores between genes are typically used in microarray analysis as evidence of coexpression.

PCC scores can also be used to predict direct targets of transcription factors, but the correlation of direct targets of transcription factors is delayed in time. The PCC calculated at various time lags is called the cross-correlation. This accounting for time-lag is due to the fact that translation of the transcription factor mRNA must occur before the changes in expression of a direct target can be measured. A time course microarray was used to predict direct targets of transcription factors using cross-correlation (Kato and Tsunoda, 2001). The cross-correlation function was calculated for the time-series expression data for transcription factors and potential targets in the yeast genome during the mitotic cell cycle. Statistically significant correlation results were used to screen for direct targets, and a significant number of predicted targets were confirmed using independent data sets.

The inverse of correlation (negative correlation scores, or anti-correlation) can be used to predict repressive effects. While the mechanism of miRNA is different than that of a transcription factor, the effect on gene expression is similar to that of a repressing transcription factor. Increases in expression levels of repressors and miRNA both cause decreases in gene expression of direct targets, the primary difference being that repressors act with a time lag, while miRNA acts contemporaneously on its targets with an intermediate step of translation. Direct targets of miRNA transcripts were predicted using anti-correlation with no time lag (Liu et al., 2010).

5.1.2: A *C. elegans* Embryonic Time Course Microarray Data Set

Baugh, et al. produced a set of wild-type time course microarray measurements of transcript levels during *C. elegans* embryogenesis. (Baugh et al., 2003) These measurements covered the first quarter of embryogenesis with precisely staged embryos up to the 190-cell stage. By this stage of development, not all cell fates are determined, but greater than 85% of all cells will have descendants which share the same primary fate. The resolution of the time course was approximately every 20 minutes, such that there are approximately two samples per cell division cycle. Thus, this time course microarray data set should cover most cell-fate related events in embryonic patterning.

A second time course microarray data set was subsequently prepared, covering the same time period at similar intervals, with improved replicate coverage and using publically available Affymetrix GeneChips. (Baugh et al., 2005) This second microarray data set is used in this analysis, and covers the expected period of *tbx-2* expression and likely early function, despite not covering all of embryogenesis.

5.1.3: Summary of Analysis

Here, a *C. elegans* embryonic time course microarray data set (Baugh et al., 2005) is used to map potential direct targets of TBX-2 by combining the methods of time-lagged correlation and anti-correlation, called time-lagged anti-correlation (TiLAC).

5.2: Materials and Methods

Cross-correlation was calculated (Kato and Tsunoda, 2001) using the *ccf* function in Bioconductor (Gentleman et al., 2004) on expression values at each time point for *tbx-2* and each gene in the embryonic time course microarray data set for 0-minutes lag, 20-minutes lag, and 40-minutes lag. Two probes in the microarray map to the genomic region for *tbx-2*, one in a coding region (188633_at), and one in the 3'-UTR (174869_at). Both have substantially similar expression patterns ($r > 0.99$), but the coding region probe showed higher expression levels. The coding region probe for *tbx-2* was used for all subsequent analysis. Plots were generated using the *gplot* function for the cross-correlation function, and for the expression levels for *tbx-2* and each gene tested. Correlation coefficients were evaluated for statistical significance (Kato and Tsunoda, 2001) and multiple test correction was performed on the results (Benjamini, 1995).

5.3: Results

5.3.1: Expression Profiles for *tbx-2*

The *tbx-2* expression profile was calculated from the embryonic time course microarray data set (Baugh et al., 2005) and graphed (Figure 15). Two probes in the microarray map to the genomic region for *tbx-2*, one in a coding region, and one in the 3'-UTR. Both had substantially similar expression patterns, but the coding region probe showed higher expression as might be expected due to 3' RNA degradation in microarray preparation. The coding region probe was used for subsequent TiLAC analysis. The peak beginning around the 100-cell stage corresponds with known *tbx-2* embryonic expression patterns.

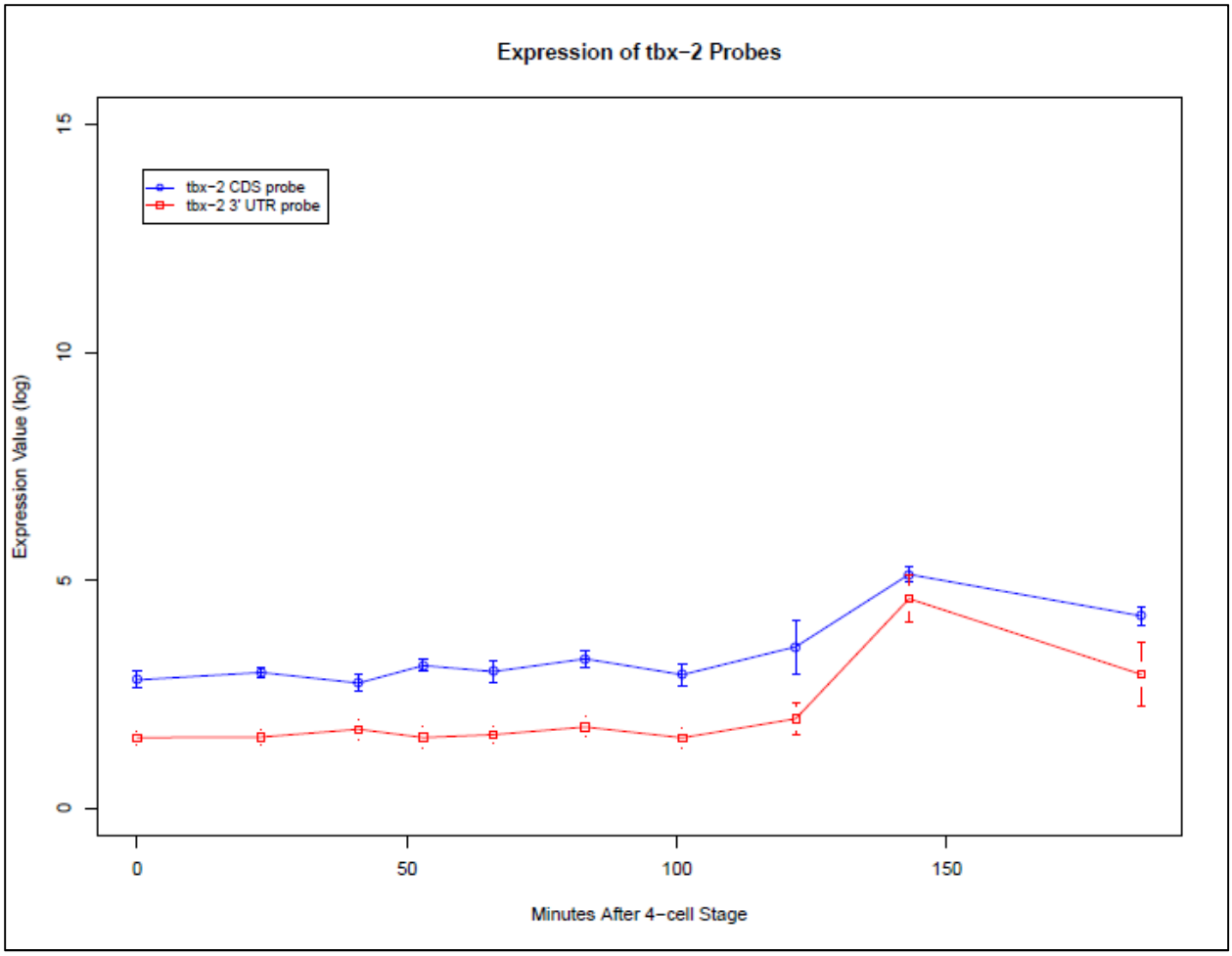

Figure 15: tbx-2 Expression Profile

Expression levels of the two *tbx-2* probes from the embryonic time course microarray data set (Baugh et al., 2005). The blue line represents 188633_at, a probe mapping to the coding region, while the red line represents 174869_at, a probe mapping to the 3'-UTR. Overall levels remain relatively flat until the 100-cell stage, at which point *tbx-2* levels increase.

5.3.2: Analysis of the embryonic time-course microarray data for correlation of expression with tbx-2

The distribution of correlation with *tbx-2* at no lag, at 20-minutes lag, and 40-minutes lag was plotted (Figure 16). Cross-correlation values depend on the expression level similarity at a particular time point, and on the number of time points being compared (Kato and Tsunoda, 2001). Cross-correlation of the Baugh, et al., data set consists of 10 time points at no lag, of 9 data points at 20-minutes lag, and of 8 time points at 40-minutes lag. Correlation values approaching 1 represent high relatedness between the expression profiles for positive correlation. Correlation values approaching -1 represent high relatedness between expression profiles but change in the opposite direction. Correlation near 0 represents no relationship between the expression profiles. Statistical significance depends on the value of correlation as well as on the number of time points (Kato and Tsunoda, 2001).

The distribution of all correlation values for *tbx-2* is shown in Figure 16. The no lag distribution shows the most number of probes with high-relatedness (probes with PCC scores approaching 1 or -1), while the 40-minute lag has the fewest (Figure 16). Statistical significance for anti-correlating probes shows a similar trend due to reduced number of data points available due to time-shift (Figure 17). At no lag, 3903 probes anti-correlate ($p < 0.01$). At 20-minutes time lag, 974 probes anti-correlate ($p < 0.01$). At 40-minutes lag, no probes anti-correlate ($p < 0.01$). After multiple test correction, only 1 probe correlation is significant: the probe representing *hsp-12.6* with a correlation score of -0.97 at no lag.

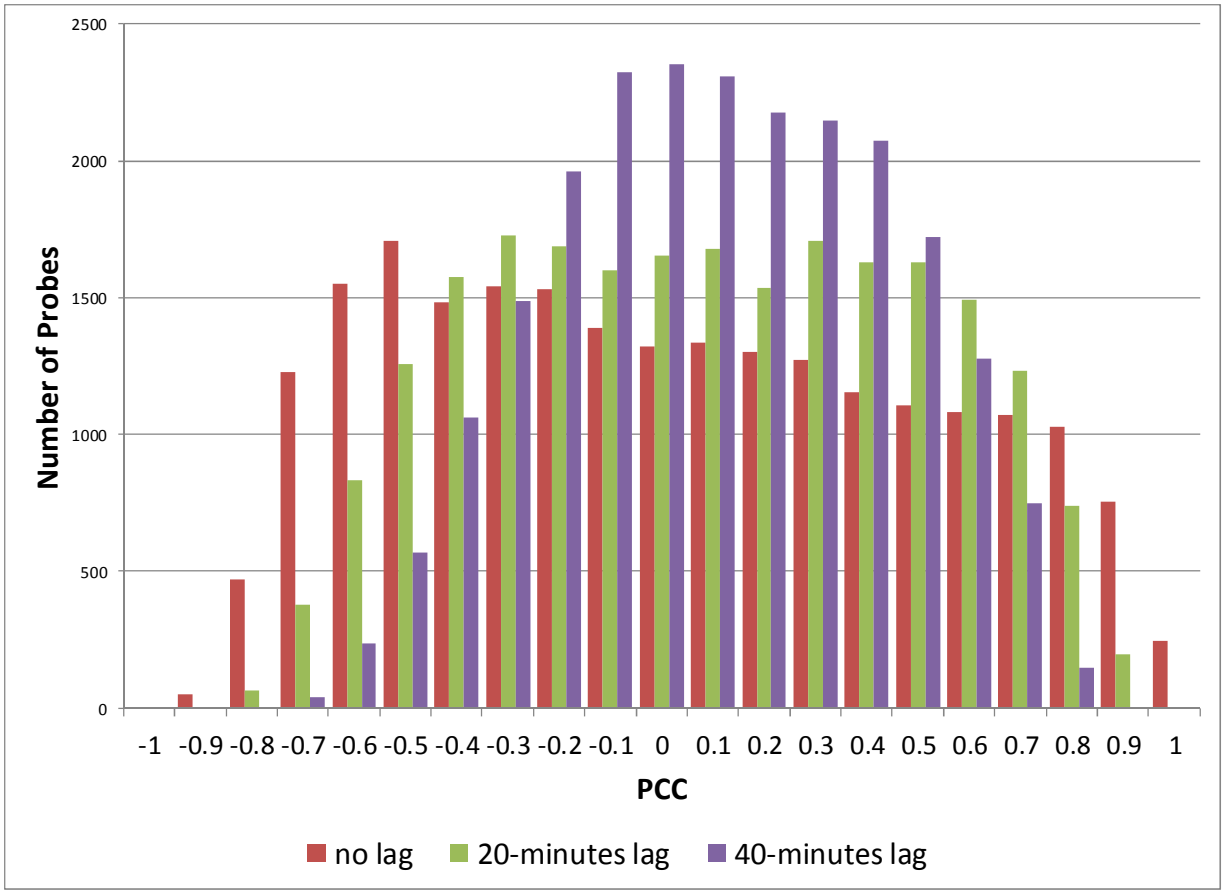

Figure 16: Distribution of tbx-2 Correlation
The complete distribution of PCC scores for *tbx-2* compared to each other gene in the data set is shown at various lag times.

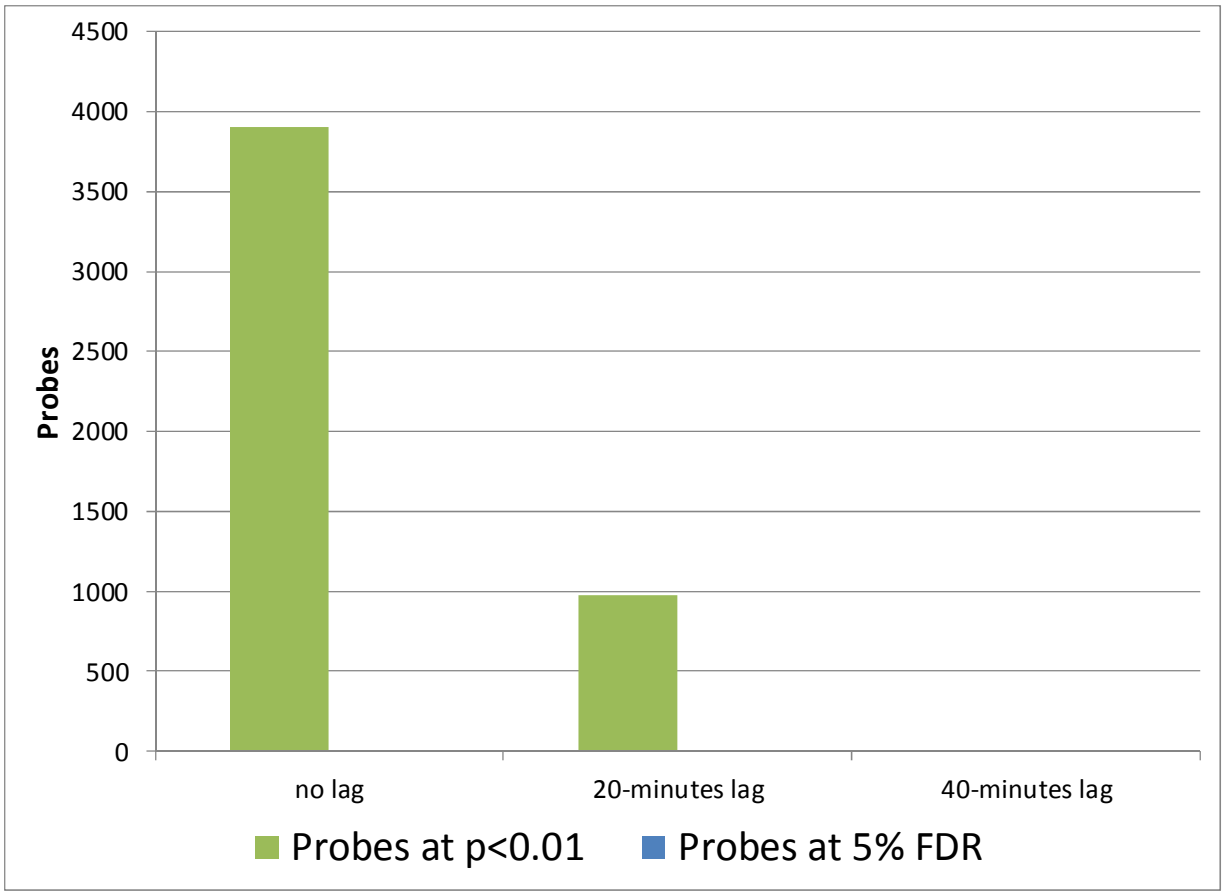

Figure 17: Anti-correlation with tbx-2

The numbers of probes which anti-correlate with *tbx-2* (at $p < 0.01$ and 5% FDR after multiple test correction are shown). At 5% FDR, only one probe correlates with *tbx-2,* corresponding to *hsp-12.3* at no lag time.

5.4: Discussion

5.4.1: TBX-2 May Be Active Earlier Than Expected From Previous Expression Analysis

The increase in expression for *tbx-2* beginning at the 100-cell stage on the embryonic time course microarray data set corresponds to previously described *tbx-2* expression in embryos. However, the non-zero levels of tbx-2 RNA as soon as the 4-cell stage may indicate an earlier role for tbx-2. Early products are likely to be due to parental contribution, and the changing levels near the 100-cell stage likely represent zygotic transcription.

5.4.2: Time-lagged anti-correlation (TiLAC) Analysis Results Should Include Potential Direct Targets of TBX-2

Genes downstream of *tbx-2* can be identified by measuring differential expression in the mutant as compared to wild type, such as in a two-state microarray experiment, but this measurement can only result in two states: *change*, or *no change*. A time-course microarray experiment results in expression values over time (an expression profile) which provides additional information which can be utilized. The expression profile of a transcription factor can be related to the expression profile of its targets. Working under the hypothesis that *tbx-2* is acting as a repressor, directly repressed targets may show expression level changes opposite to those of *tbx-2* transcription levels, and the response should demonstrate a time lag accounting for transcription and translation. Too long a time lag, or no time lag, and the target is unlikely to be a direct target of a transcription factor.

A well-characterized cascade of *C. elegans* transcription factors is analyzed in the context of an embryonic time-course microarray experiment, and relevant lag times of 20 to 40

minutes are seen (Baugh et al., 2005). Although some lag time is an important aspect of TiLAC analysis, the resolution of the data set is limited to 20-minutes, and so a lag of 10-minutes may be masked. Thus, no-lag correlation is considered as well.

A significant number of probes correlate with *tbx-2* at statistically significant levels (Figure 17). Each of these probes represents a gene which is expressed in an inverse relationship to *tbx-2* and with a time lag of 0 to 20 minutes. These represent a pool of potential direct targets of TBX-2.

5.4.3: The Results of Time-Lagged Anti-correlation (TiLAC) also Contain False Positives

While the results of the TiLAC analysis represent a pool of potential direct targets of TBX-2, the results also contain genes which are almost certainly not direct targets of TBX-2. Since *tbx-2* demonstrates transient embryonic expression, other transcription factors with similar transient expression patterns might produce similar results. Specifically any transcription factor with significant positive correlation at no time lag to *tbx-2*, and which acts as a repressor could result in some of the same predictions. Thus, additional criteria would be useful to add to this analysis in order to improve the accuracy of the predictions for direct targets of TBX-2, such as transcription factor binding site predictions, or differential expression data.

5.4.4: The Results Lack High Statistical Significance After Multiple Test Correction

While useful for screening purposes (Kato and Tsunoda, 2001), the limited number of overlapping data points (n = 10, 9, or 8 depending on time-lag) limits the potential statistical significance of the data set. A data set with higher density or larger number of time points would dramatically improve statistical significance. Since the transient expression of *tbx-2* is biased towards the end of the data set, even two additional data points in the data set at the end would improve the power of the analysis for 20-minute lag and 40-minute lag and the trends of decreasing statistical significance would be significantly ameliorated. In its current form, the data set can also be useful for transcription factors with key expression earlier in the data set, as well as for analysis of upstream factors for *tbx-2*.

5.4.5: TiLAC Results May Predict a Limited Class of Targets

Predicting targets of transcription factors using correlation (or anti-correlation) has significant limitations. In a simple network, the relationship between a gene and its target can correlate strongly (Kato and Tsunoda, 2001; Alon, 2007). But gene regulatory networks are vast in scope and involve hundreds of genes. Even simple three-gene or four-gene systems can display remarkable complexity of expression patterns and regulatory control (Alon, 2007). Since *tbx-2* is known to be downstream of other genes and has auto regulatory capabilities (A. Milton and P. Okkema, personal communication), expression patterns of some direct targets may not follow simple patterns. Thus, the assumption in TiLAC that direct targets will correlate may only apply to a limited class of direct targets, and thus may significantly under predict potential direct targets.

5.4.6: Combination With Other Prediction Methods Should Enhance the Likelihood of Finding Direct Targets

While the TiLAC method produces several thousand potential direct targets, this number is still too many for biological testing in most circumstances. Combining the method with additional, independently-sourced predictions of direct targets should help filter out false positives unique to this method and improve the predictive power of the analysis.

CHAPTER 6: COMBINING METHODS FOR PREDICTING DIRECT TARGETS OF TBX-2

6.1: Introduction

Several independent methods for determining direct targets of transcription factors have been presented in the preceding chapters. Each method has strengths and weaknesses, and each method is based on completely independent data sets and assumptions.

Microarray analysis of mutant vs. wild type gene expression results in potential downstream targets, but is unable to differentiate between direct and indirect targets. The two condition microarray experiment lacks detailed timing information, presents no information on time-dependent expression profiles, and is silent about whether differentially expressed genes have potential binding sites in the regulatory region.

Prediction of transcription factor binding sites, in this case with the TBXCore model, results in predictions for a broad family of T-box transcription factors rather than specifically for TBX-2. In addition, the prediction of binding sites relies only on sequence characteristics in a subset of regulatory regions in the genome, and results in a large number of predicted binding sites. These binding site predictions can be refined by including analysis about conservation and position in order to increase the likely biological relevance of a predicted binding site. These methods results in a significant number of false positive predictions, but can be used to exclude potential direct targets with greater certainty. Nevertheless, these binding site predictions present no information about expression patterns or timing.

Time-course microarray data analysis, as presented here, gives better resolution than a two-state microarray but only covers wild type expression patterns. Predictions of direct targets of TBX-2 can be made based on expression profiles combined with time-lag, but these predictions contain results from transcription factors with similar timing and behavior. Furthermore, these predictions are independent of evidence of binding sites in regulatory regions, or of differential expression in the mutant.

Each of these methods presents a strong foundation for finding direct targets. When used alone, they result in large numbers of predictions containing significant numbers of false positives. However, since each method does not share the same assumptions, it should not share the same set of false positive predictions. Thus the results from each method, when combined, provide a pool of candidate with multiple, independent factors recommending them as direct targets of *tbx-2*, resulting in a stronger, more useful, and more reliable predictor than each independently.

6.2: Materials and Methods

To combine the *tbx-2* microarray with transcription factor binding site data, the set of genes represented by the differentially expressed probes up regulated in the mutant at a false discovery rate of 5% on the *tbx-2* microarray was used. The two data sets were intersected at the gene-level, and comparisons at multiple conservation levels and proximity were made.

To combine the *tbx-2* microarray with TiLAC analysis, the set of differentially expressed probes which were up regulated in the mutant at a false discovery rate of 5% on the *tbx-2* microarray was used. This subset was then subjected to TiLAC analysis. Multiple-test correction

was then performed using Benjamini-Hochberg, and thresholds were set at a false discovery rate of 10% (Benjamini, 1995).

To combine all three sets of results, the set of genes represented by the differentially expressed probes up regulated in the mutant at a false discovery rate of 5% on the *tbx-2* microarray was used. This subset was intersected with the transcription factor binding site data sets and comparisons at multiple conservation levels and proximity were made. The probes represented by the genes in the combinations, were then subjected to TiLAC analysis. Multiple-test correction was then performed using Benjamini-Hochberg, and thresholds were set at a false discovery rate of 10% (Benjamini, 1995).

6.3: Results

6.3.1: Microarray Analysis Combined With Transcription Factor Binding Site Prediction

In order to identify direct targets of TBX-2, and working under the hypothesis that *tbx-2* is acting as a repressor, we can combine the portion of the microarray results which show up-regulation in the mutant background at a 5% false discovery rate with the results from the transcription factor binding site analysis. These results can be combined at the gene level, as an intersection of sets.

Thus, the result of the combination consists of genes which are differentially expressed, up-regulated in the mutant, and have TBXCore binding sites in the regulatory region. These results are presented at various levels of conservation and for both the complete set, and the proximate set with one or more binding sites within 1000 base pairs of the start codon (Figure 18). The combination of the data sets results in a significant reduction of potential targets over

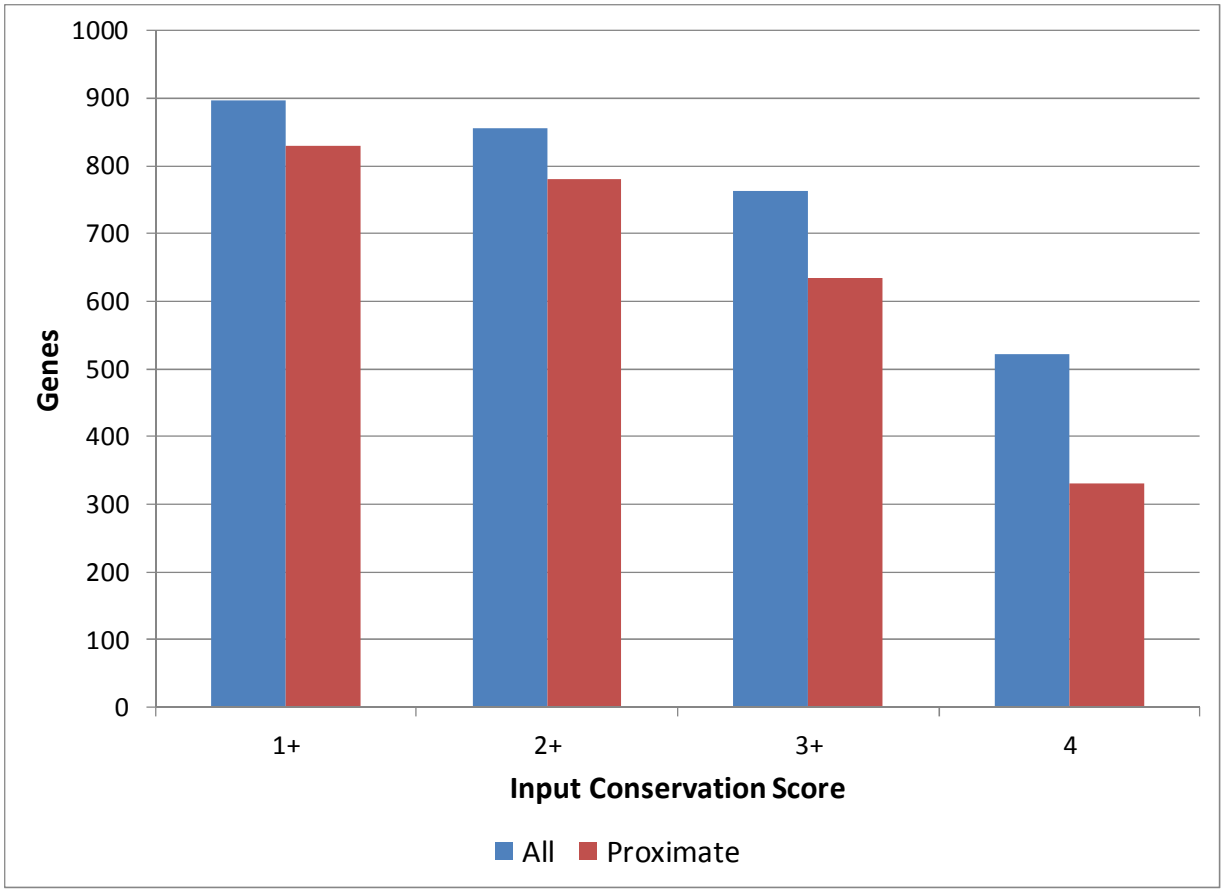

Figure 18: Microarray Results Combined with Binding Site Predictions
Quantities of differentially expressed genes, up-regulated in the mutant, which have TBXCore binding sites in the regulatory region, are shown. Proximate sites reflect genes which have one or more binding sites within 1000bp of the start codon.

either individual results set. The potential targets are reduced by 10-60% over the microarray alone, and 94%-98% over the TBXCore model alone, depending on the level of conservation. This reduction in number of potential direct targets is expected to come with a quality increase in those predicted targets due to having two independent sources of predictions.

6.3.2: Microarray Analysis Combined With Correlation Analysis

The combination of the microarray results can also be combined with the correlation analysis at the probe level. With the goal of identifying direct targets of TBX-*2* acting as a repressor, the portion of the microarray results which show up regulation in the mutant background at a 5% false discovery rate were combined with the results from the correlation analysis. These results were be combined at the probe level, as an intersection of sets, and then mapped to genes. The combination of these two sets results in a very small set of potential direct targets: fewer than 100 targets at no lag, and fewer than 10 targets at 20-minutes lag, and no targets at 40-minutes lag ($p < 0.01$) (Figure 19).

The statistical power of correlation analysis can be increased by pre-filtering the list of genes to compare with expression patterns of *tbx-2* by the results of the microarray. This results in fewer comparisons during correlation analysis (~1200 vs ~20,000) and a correspondingly smaller multiple-test correction factor. While multiple test correction on the correlation data alone resulted in no results, this method results in four rigorously statistically significant potential direct targets at a false discovery rate of less than 10%. These targets – *hsp-12.3* (Figure 20), *lea-1* (Figure 21), *K08H10.2* (Figure 22), and *M03E7.2* (Figure 23) – represent a small, rigorous set for biological testing.

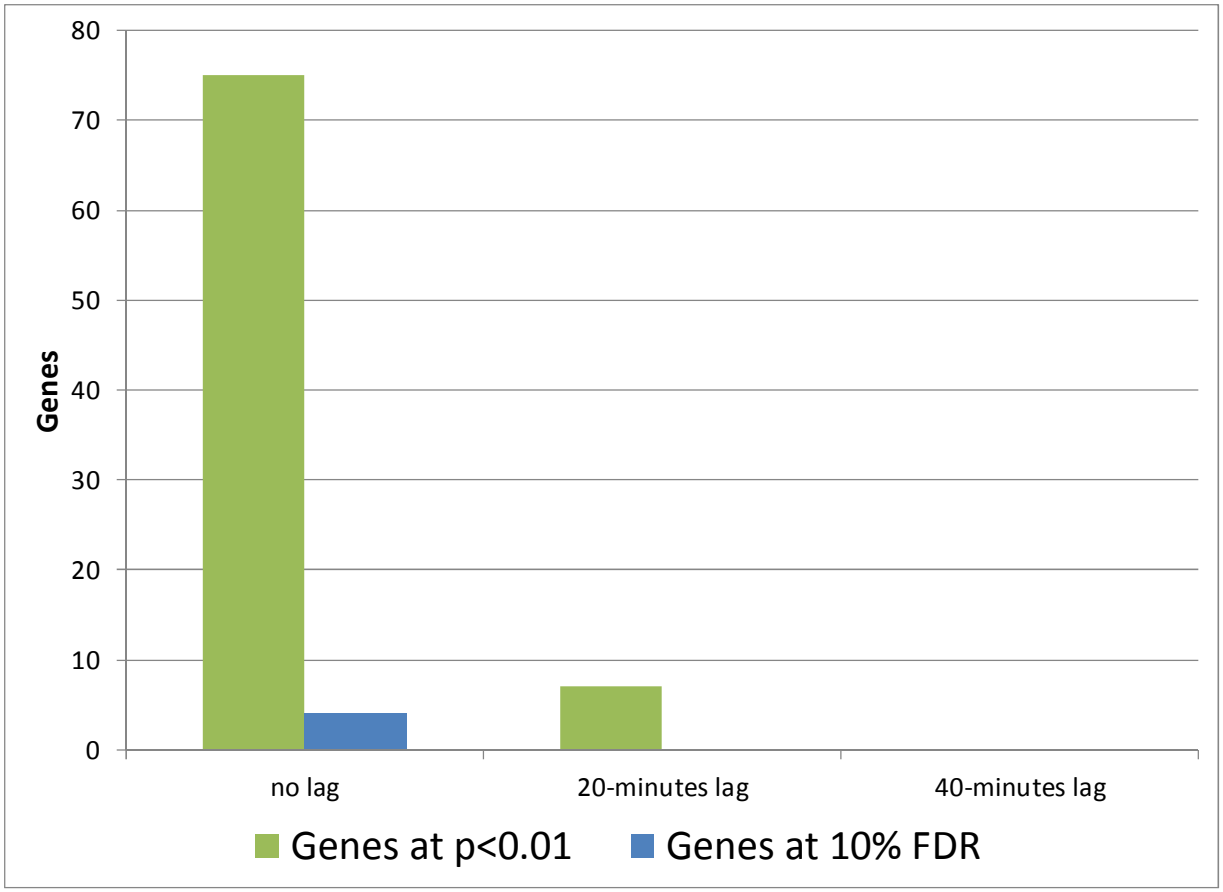

Figure 19: Microarray Results Combined with Correlation Analysis
Quantities of differentially expressed genes, up-regulated in the mutant, which anti-correlate with tbx-2 (at $p < 0.01$ or at a multiple test corrected FDR of 5%) are shown.

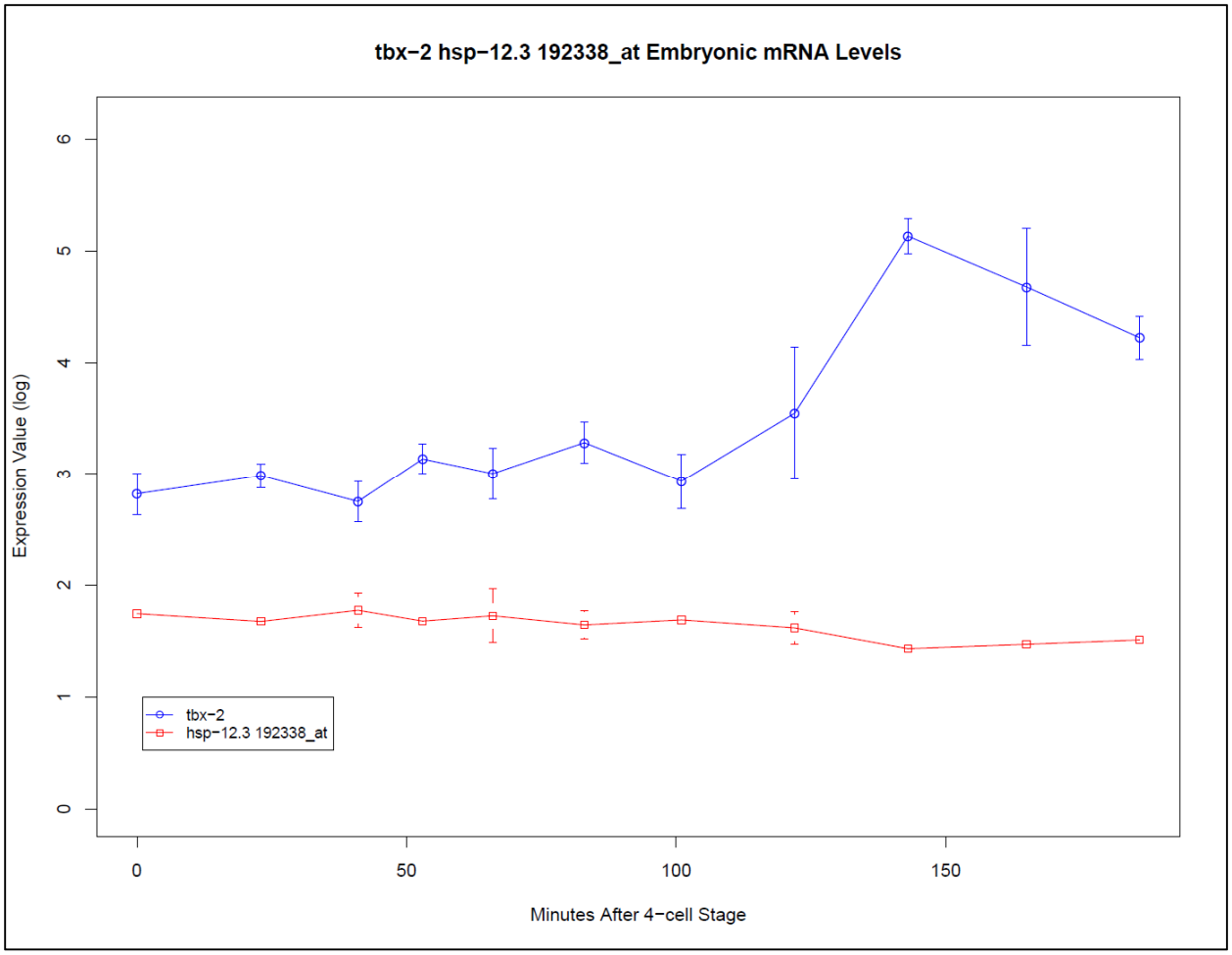

Figure 20: Expression Profile of *tbx-2* Compared to *hsp-12.3*

Expression levels of *tbx-2* compared to *hsp-12.3* ($r < -0.97$) from the embryonic time course microarray data set (Baugh et al., 2005). The blue line represents *tbx-2*, and the red line represents *hsp-12.3*.

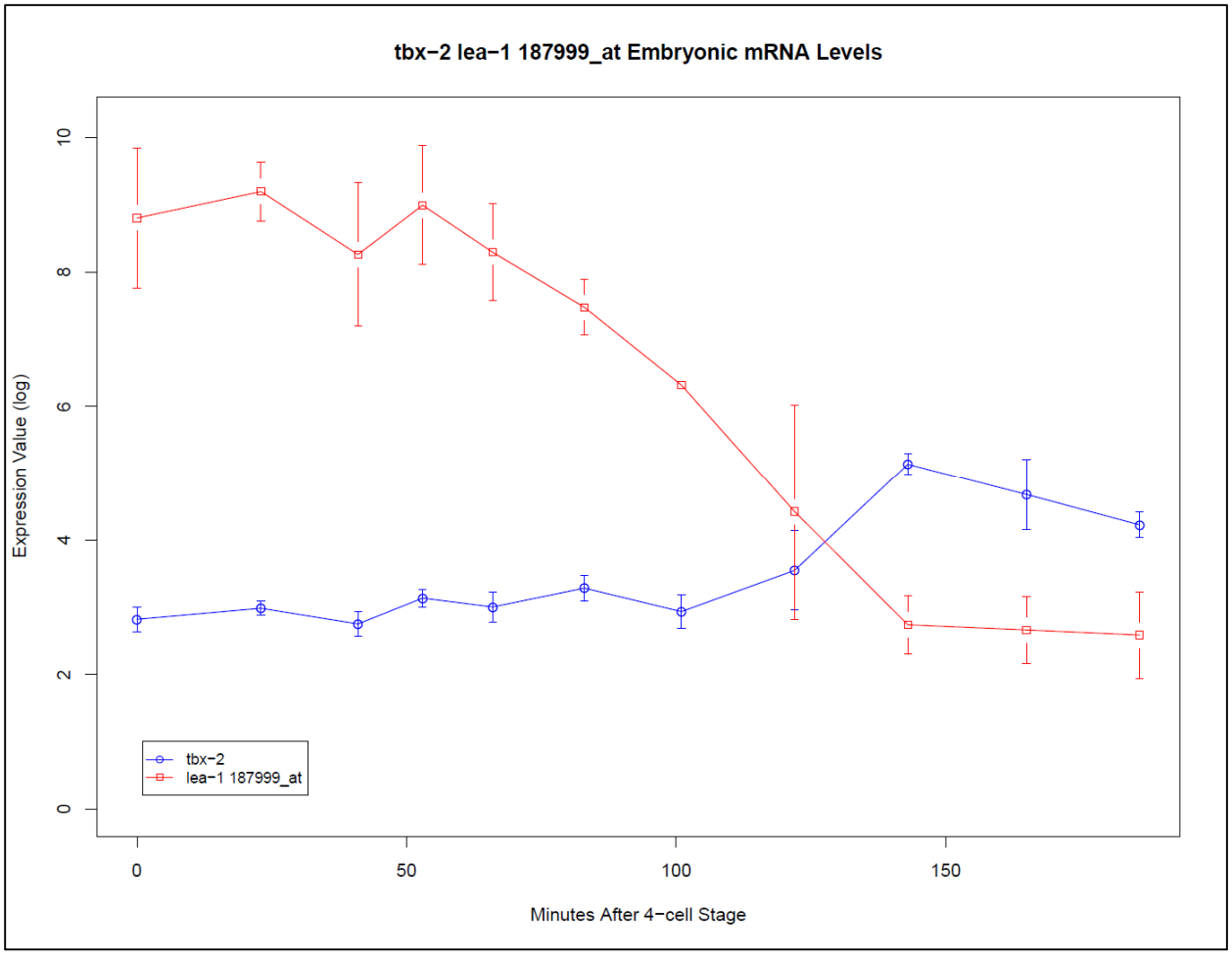

Figure 21: Expression Profile of *tbx-2* Compared to *lea-1*

Expression levels of *tbx-2* compared to lea-1 ($r < -0.91$) from the embryonic time course microarray data set (Baugh et al., 2005). The blue line represents *tbx-2*, and the red line represents *lea-1*.

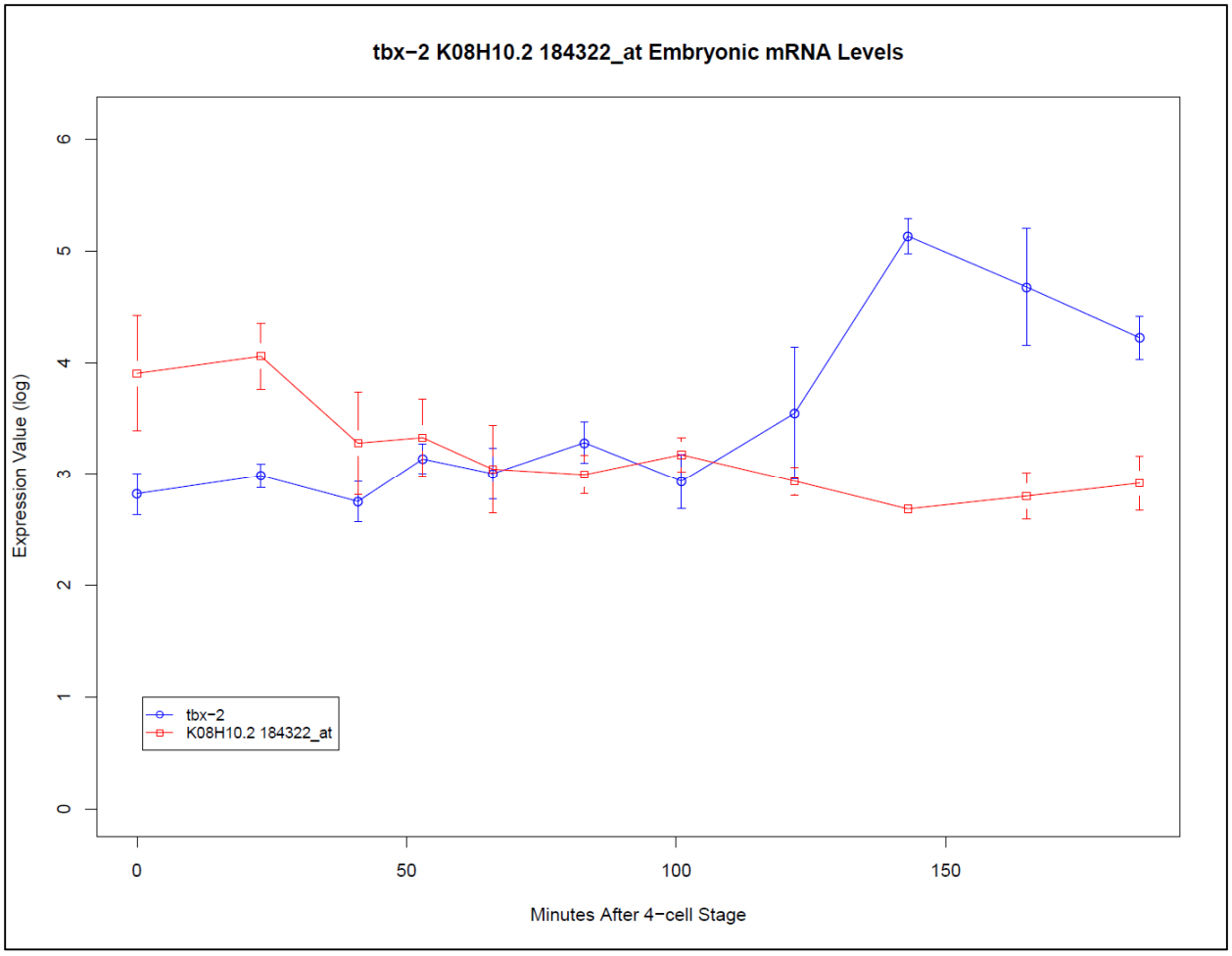

Figure 22: Expression Profile of *tbx-2* Compared to *K08H10.2*

Expression levels of *tbx-2* compared to *K08H10.2* (r < -0.91) from the embryonic time course microarray data set (Baugh et al., 2005). The blue line represents *tbx-2*, and the red line represents *K08H10.2*.

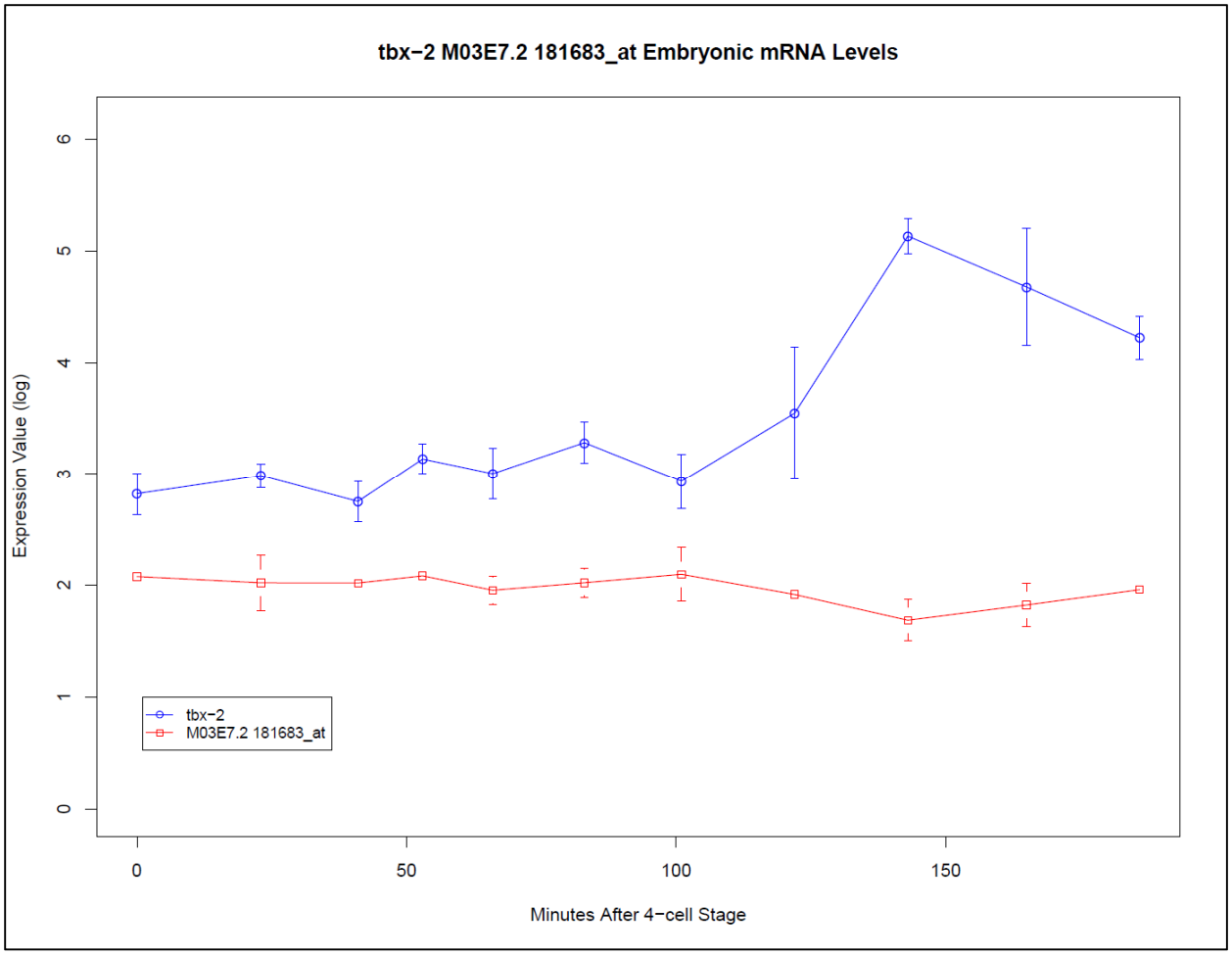

Figure 23: Expression Profile of *tbx-2* Compared to *M03E7.2*

Expression levels of *tbx-2* compared to *M03E7.2* (r < -0.89) from the embryonic time course microarray data set (Baugh et al., 2005). The blue line represents *tbx-2*, and the red line represents *M03E7.2*.

6.3.3: Microarray Analysis, Transcription Factor Binding Site Prediction, and Correlation Analysis

The microarray analysis, transcription factor binding site prediction, and correlation analysis can be combined to further reduce the set of potential direct targets for biological testing. Compared to the combination of microarray analysis and correlation analysis (Section 6.3.2: above), the addition of the transcription factor binding site prediction has significant effect, reducing predicted targets by 20% to 76% at no time-lag, depending on the level of conservation and location of binding sites (Figure 24). At 20-minutes lag, the predicted targets are reduced by 20%-50%, though they are few in number (Figure 25). With the three-way data set combination, at 40-minutes lag there are no predicted targets (data not shown).

As above, the statistical power of correlation analysis can be increased by pre-filtering the list of genes to compare with expression patterns of *tbx-2* by the results of the microarray and the binding site analysis. Prefiltering using the best set from the microarray analysis combined with the binding site analysis (331 genes, at a proximate input conservation score of 4) results in a smaller multiple test correction factor. Nevertheless, at this more rigorous multiple-test corrected threshold, no targets remain (data not shown).

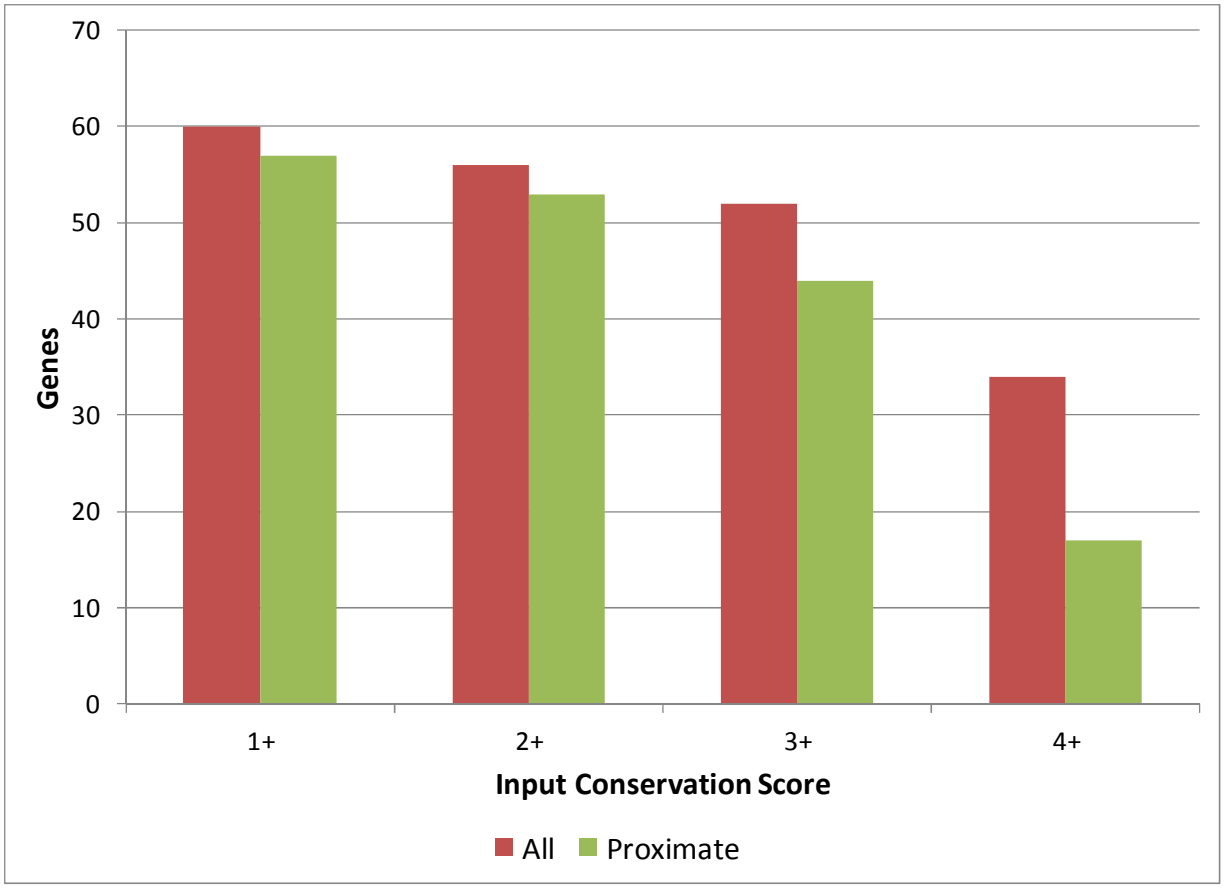

Figure 24: Microarray Results Combined with Binding Site Predictions and Correlation Analysis (no lag)

Quantities of differentially expressed genes, up-regulated in the mutant, which anti-correlate with tbx-2 (at $p < 0.01$ or at a multiple test corrected FDR of 5%) at no time lag and which have TBXCore binding sites in the regulatory region are shown. Proximate input conserved genes have one or more binding sites within 1000bp of the start codon.

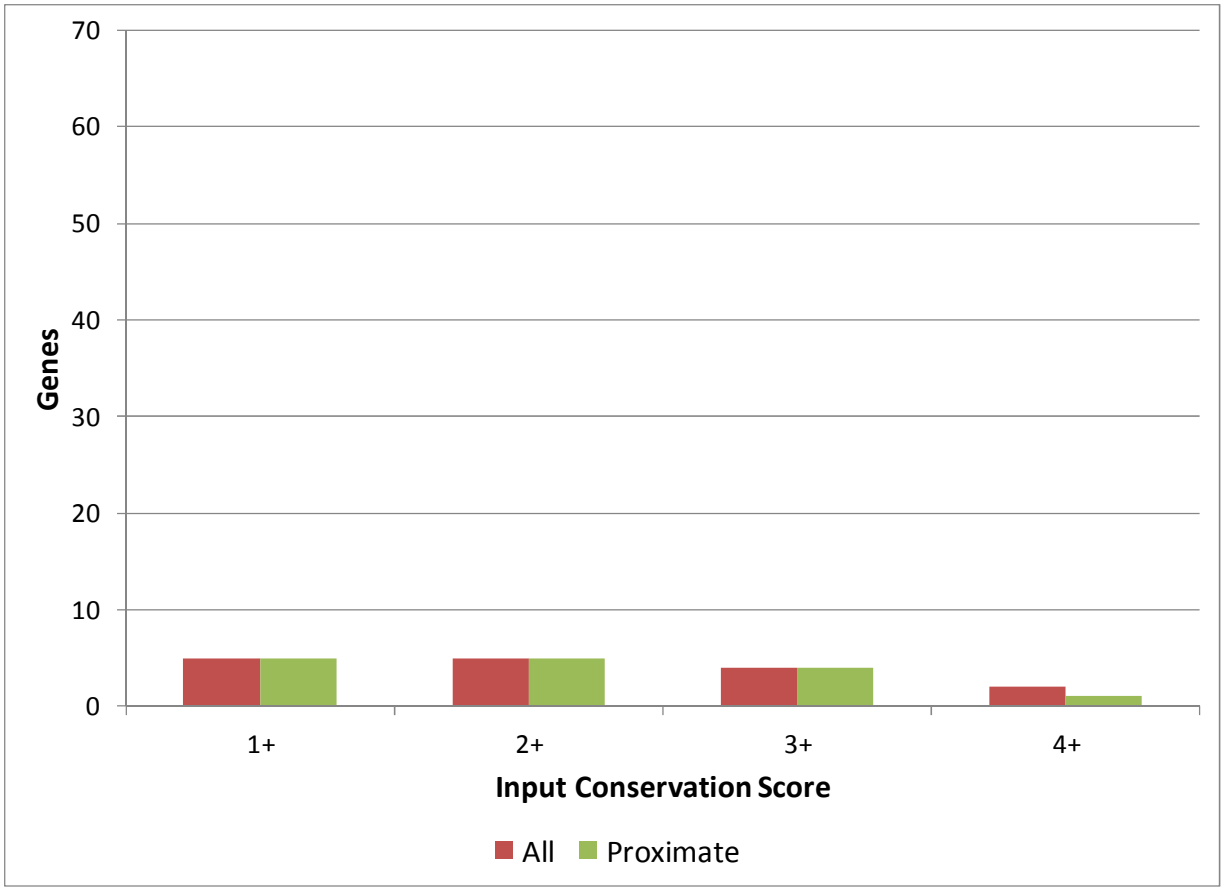

Figure 25: Microarray Results Combined with Binding Site Predictions and Correlation Analysis (20-minutes lag)

Quantities of differentially expressed genes, up-regulated in the mutant, which anti-correlate with tbx-2 (at $p < 0.01$ or at a multiple test corrected FDR of 5%) at 20-minutes time lag and which have TBXCore binding sites in the regulatory region are shown. Proximate input conserved genes have one or more binding sites within 1000bp of the start codon.

6.3.4: Overview of Effects of Combined Analysis and Application to Biological Testing

Starting with the microarray analysis as a baseline, 1179 genes are predicted as potential downstream targets of TBX-2. When the results of the TBXCore model (at a threshold of 0.80, independent of conservation) are combined, the number of potential direct targets is reduced to 939. The addition of Input Conservation criteria to improve the quality of binding site prediction (Input Conservation score of 3+, proximate sites only) reduces the number of potential direct targets to 634. Finally, restricting the criteria in the TiLAC analysis to a non-zero time-lag, and anti-correlation with $p < 0.01$ reduces the number of potential direct targets to 4 (Figure 26). These four targets – *mxl-3* (Figure 27), *nex-1* (Figure 28), *F53F4.13* (Figure 29), and *Y60C6A.1* (Figure 30) – represent a small, high-quality testable set of potential direct targets predicted by multiple, independent methods.

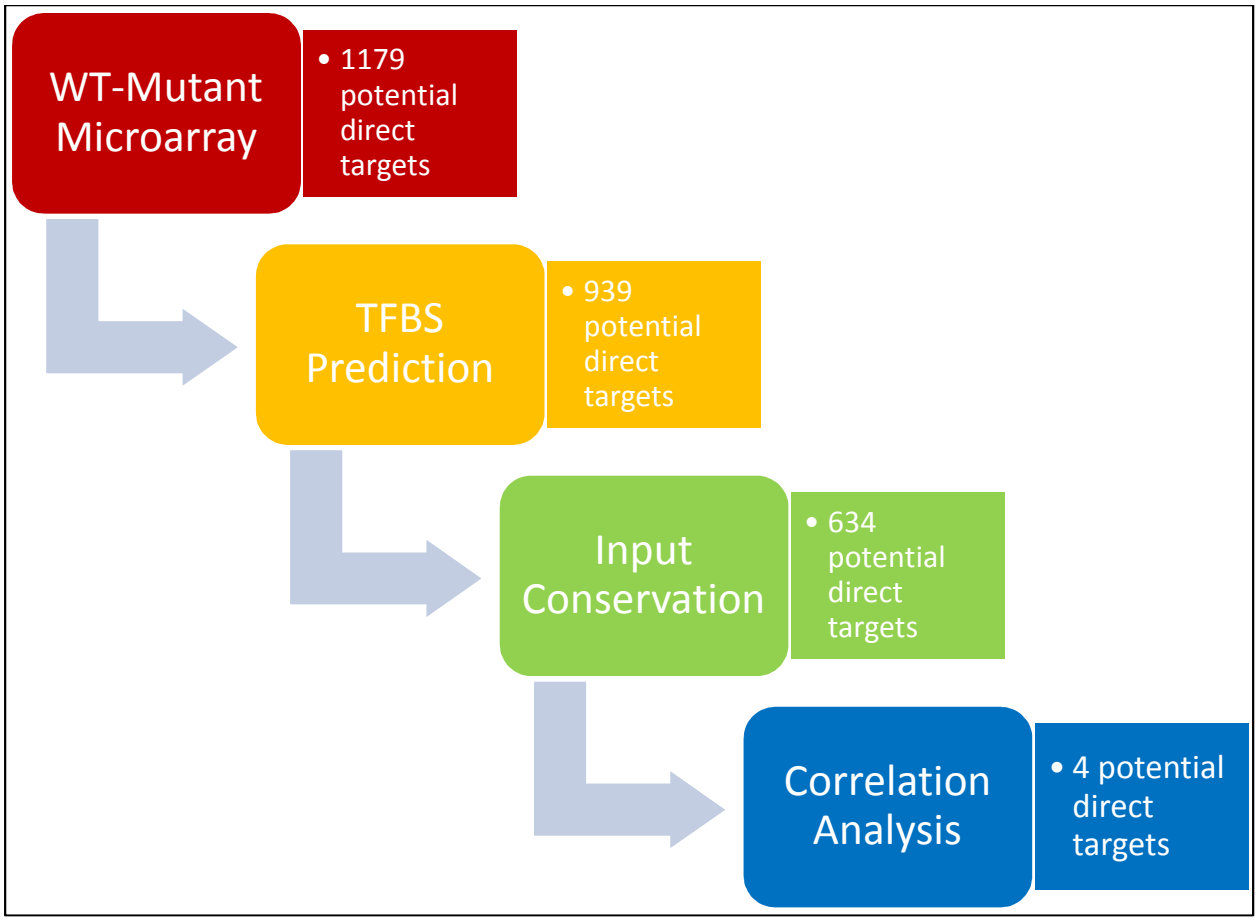

Figure 26: Overview of Reduction of Potential Direct Targets

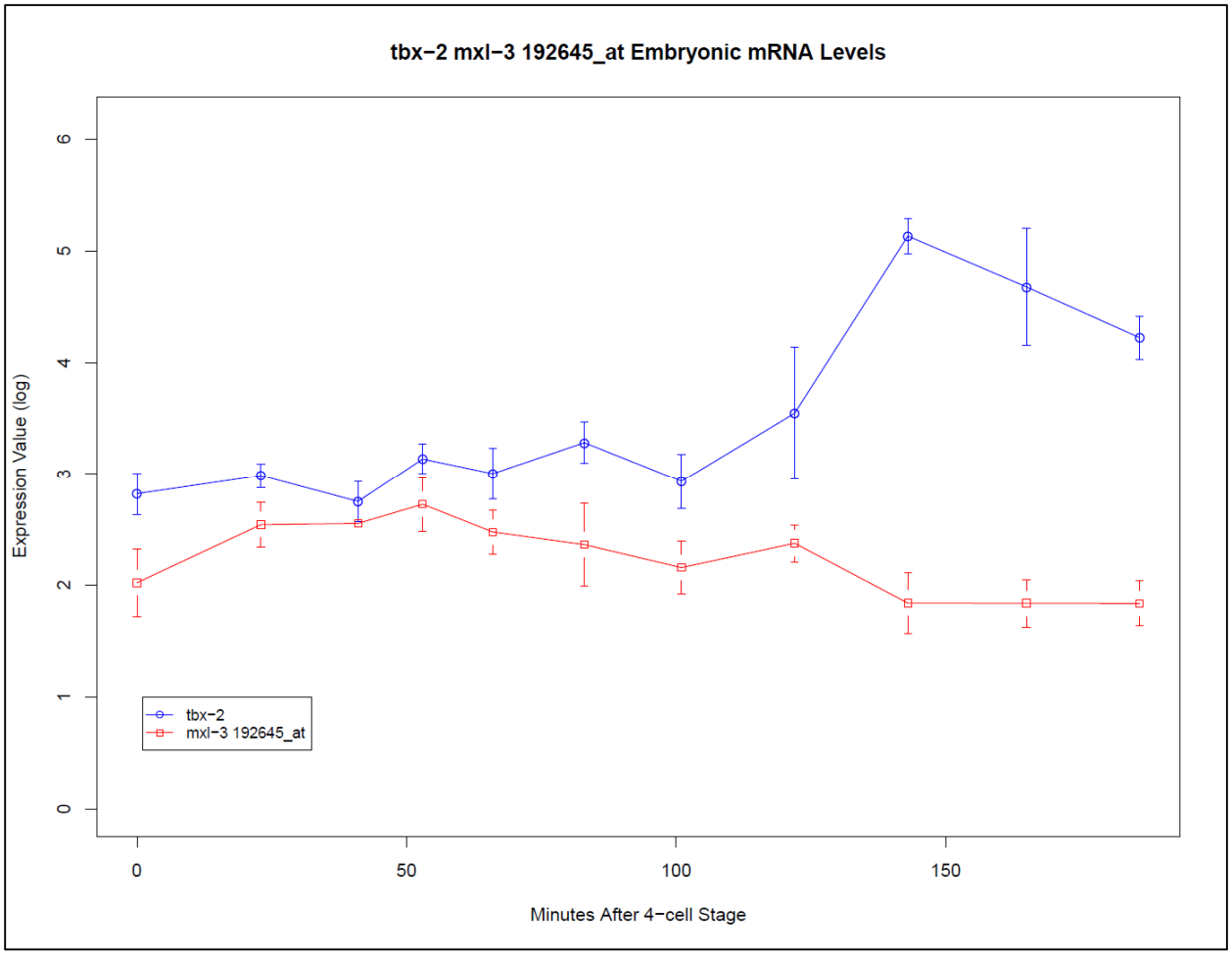

Figure 27: Expression Profile of *tbx-2* Compared to *mxl-3*

Expression levels of *tbx-2* compared to *mxl-3* ($r < -0.78$) from the embryonic time course microarray data set (Baugh et al., 2005). The blue line represents *tbx-2*, and the red line represents *mxl-3*.

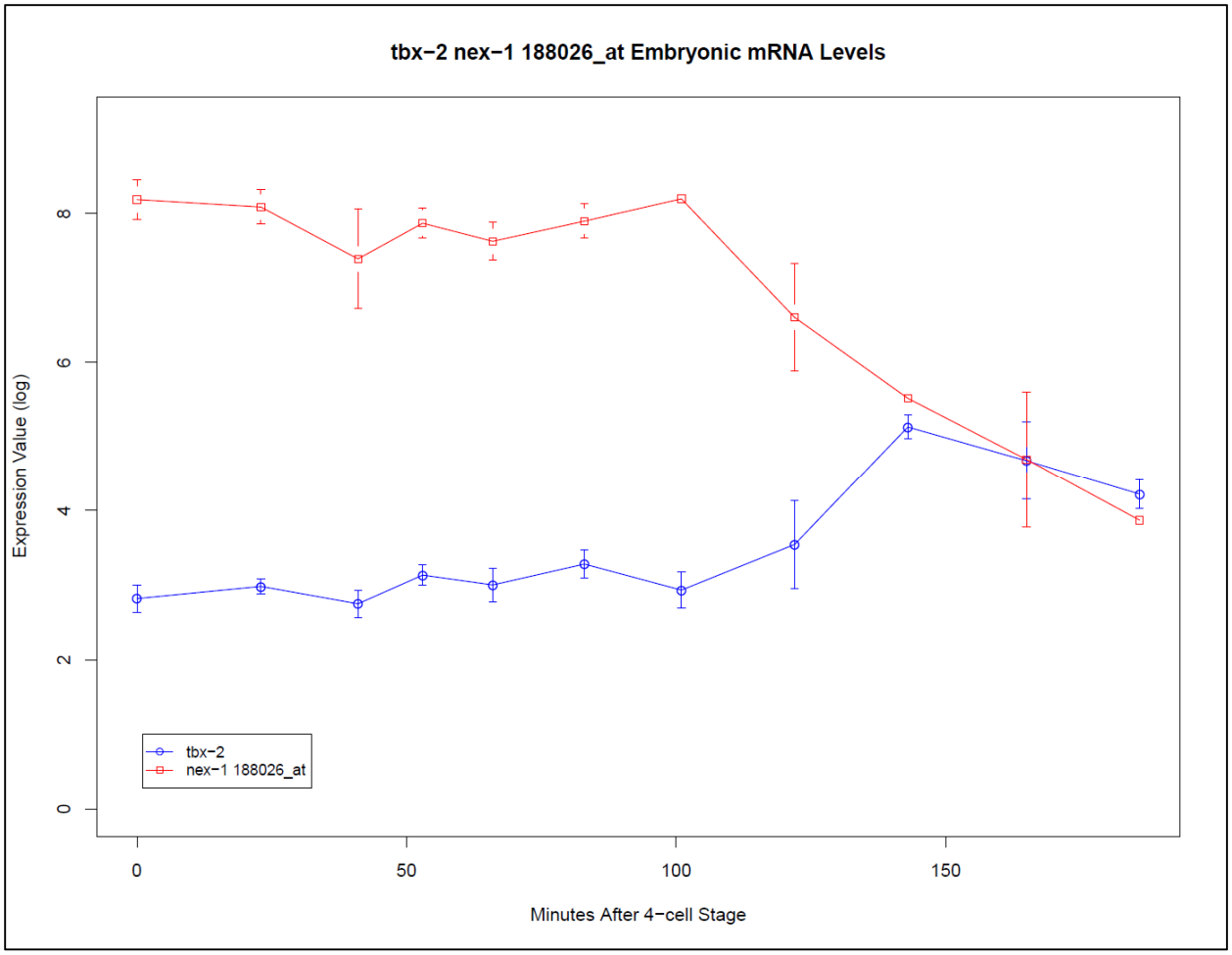

Figure 28: Expression Profile of *tbx-2* Compared to *nex-1*

Expression levels of *tbx-2* compared to *nex-1* ($r < -0.80$) from the embryonic time course microarray data set (Baugh et al., 2005). The blue line represents *tbx-2*, and the red line represents *nex-1*.

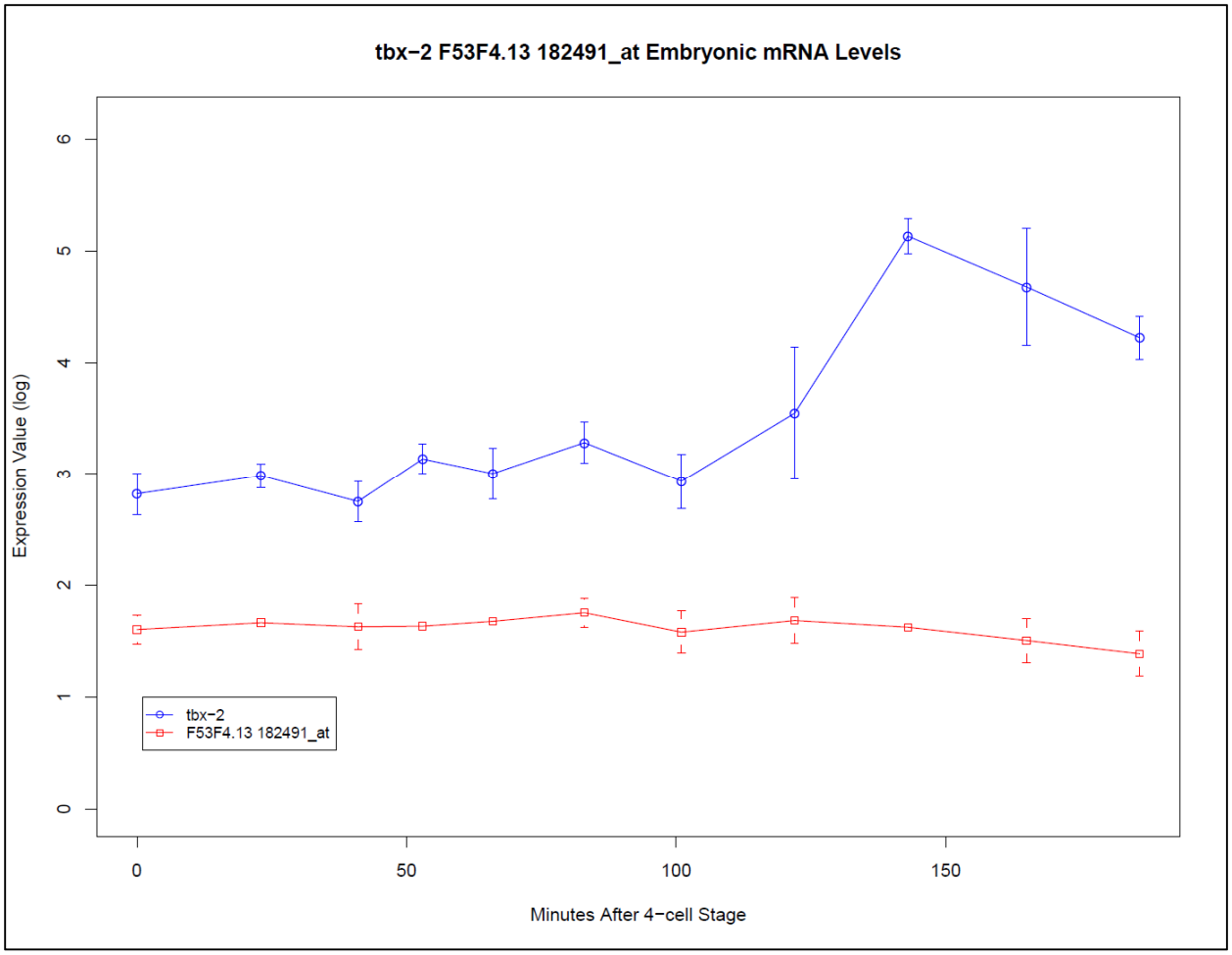

Figure 29: Expression Profile of *tbx-2* Compared to *F53F4.13*

Expression levels of *tbx-2* compared to *F53F4.13* ($r < -0.78$) from the embryonic time course microarray data set (Baugh et al., 2005). The blue line represents *tbx-2*, and the red line represents *F53F4.13*.

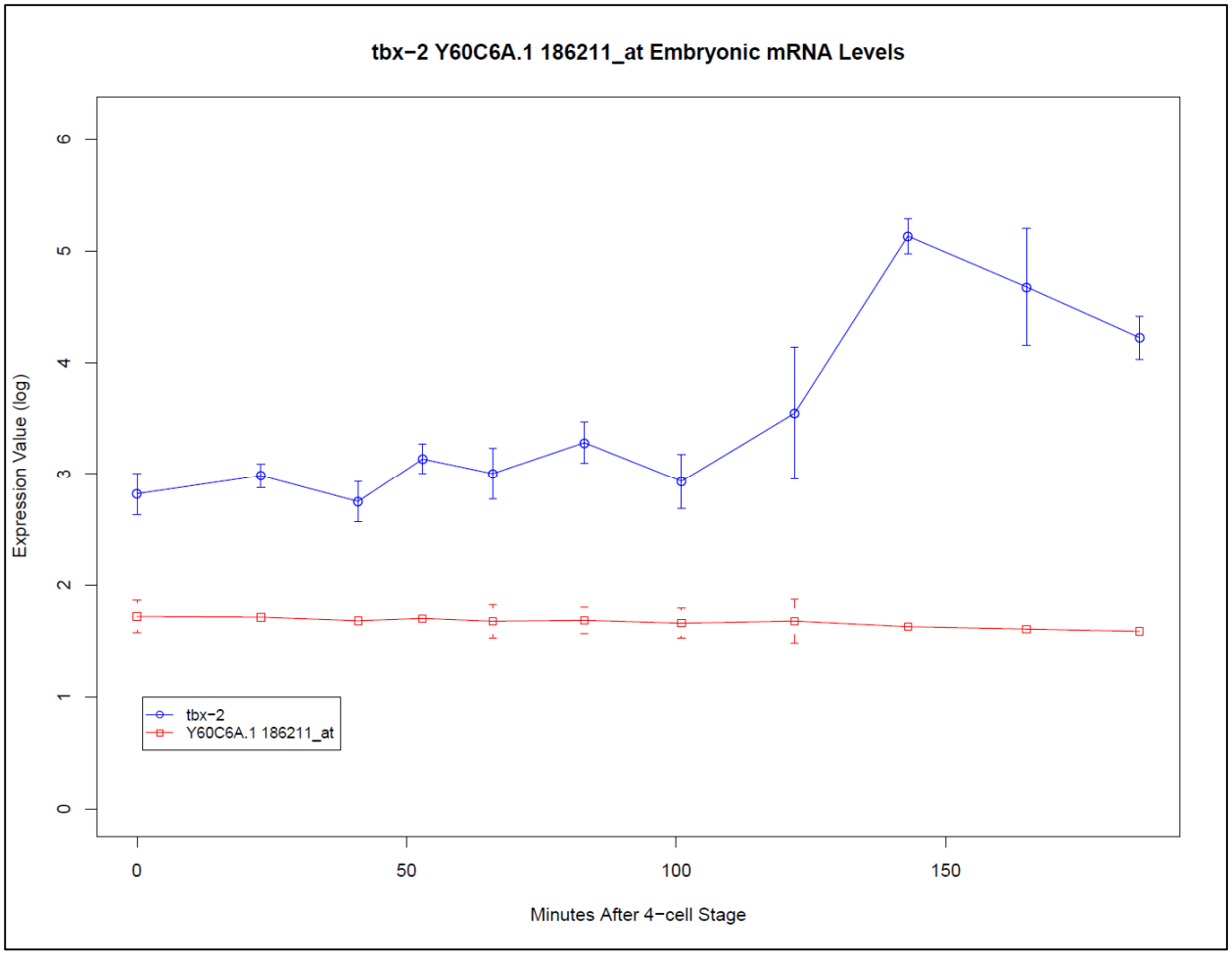

Figure 30: Expression Profile of *tbx-2* Compared to *Y60C6A.1*

Expression levels of *tbx-2* compared to *Y60C6A.1* ($r < -0.79$) from the embryonic time course microarray data set (Baugh et al., 2005). The blue line represents *tbx-2*, and the red line represents *Y60C6A.1*.

6.4: Discussion

6.4.1: Low Variance in Expression Profiles of Predicted Targets

While several predicted targets show obvious changes from high to low expression as *tbx-2* expression increases at the 100-cell stage, some of the expression profiles of predicted targets show very low variation, such as *hsp-12.3* (Figure 20), *M03E7.2* (Figure 23), *F53F4.13* (Figure 29), and *Y60C6A.1* (Figure 30). While this variation is small, the PCC scores are high indicating a strong potential for a relationship with TBX-2 ($r < -0.70$). Strong statistical significance ($p < 0.01$) indicates that it is unlikely these genes are moving in lockstep fashion and opposite direction solely due to chance. Though not shown in these graphs (as they represent wild type expression profiles only), the expression levels of each of these genes increases significantly in the mutant, and these genes have proximate input-conserved binding sites in several nematode species. The combination of these factors, rather than anti-correlation alone, is what makes them strong candidates for biological testing as direct targets of TBX-2.

6.4.2: Overview of Combined Analysis

Each individual analysis presented results in a large number of predicted targets containing potential false positive predictions. Combined analysis works to reduce the numbers of potential direct targets, and to reduce the likelihood of false positives in those predicted targets. While any intersection of multiple, non-equivalent sets will necessarily result in a smaller number of targets, the important effect of combination is to offset the effects of false predictions. Most important is the expected quality increase which comes from combining data sets containing both true positives and false positives, selected with independent criteria.

Predicted targets common to all sets should have a high likelihood of being biologically relevant direct targets of *tbx-2*, while false positive predictions unique to one method will be removed upon intersection with results from other methods. Thus, these combined methods as presented should result in a set of predictions enriched for direct targets of *tbx-2*.

APPENDICES

Appendix A

PRIMERS, PLASMIDS, AND WORM STRAINS USED IN THIS STUDY

Primers

Primer	Sequence	Description
PO910	GGAAACAGTTATGTTTGGTATA	*myo-5::GFP* fusion primer
PO939	AATATGGCAGACGATTCGATG	*myo-5* promoter amplification
PO940	AGTCGACCTGCAGGCATGCAAGCTTGGAGGGTTCATCTGTTGAG	*myo-5* promoter amplification
PO941	CTGTGTCATTTATAACCGAGG	*myo-5::GFP* fusion primer
PO1325	ACTTGGAAATGAAATAAGCTTGTACAGCACGGCGGTTTTAT	*mxl-3* promoter region for In-Fusion to pOK288.04
PO1326	TGGGTCCTTTGGCCAATCCCTTGCGGACATGGTAGAGATG	*mxl-3* promoter region for In-Fusion to pOK288.04
PO1327	ACTTGGAAATGAAATAAGCTTAATGGCCTTGTGAACCATTC	*nex-1* promoter region for In-Fusion to pOK288.04
PO1328	TGGGTCCTTTGGCCAATCCCCGATTGTAGCGTATGGGGAA	*nex-1* promoter region for In-Fusion to pOK288.04
PO1329	ACTTGGAAATGAAATAAGCTTGCAAACCGTGCTTGCTAAAT	*F53F4.13* promoter region for In-Fusion to pOK288.04
PO1330	TGGGTCCTTTGGCCAATCCCCGTCAAATGAGGAATCTACCTG	*F53F4.13* promoter region for In-Fusion to pOK288.04
PO1331	ACTTGGAAATGAAATAAGCTTGTCCCTTAGATTTGAAGCCA	*Y60C6A.1* promoter region for In-Fusion to pOK288.04
PO1332	TGGGTCCTTTGGCCAATCCCTGTCGCAGTACTTTTGGGGT	*Y60C6A.1* promoter region for In-Fusion to pOK288.04

APPENDIX A (continued)

Plasmids

Plasmids	Description
pOK246.03	MBP::TBX-2 fusion
pOK246.04	MBP::TBX-2 fusion
pOK246.05	MBP::TBX-2 fusion
pOK288.04	Pceh-28::gfp with 4xNLS

Appendix B

CREATION OF MYO-5::GFP REPORTER STRAIN

Introduction

In order to determine potential downstream targets of TBX-2, a microarray experiment was performed using Affymetrix *C. elegans* GeneChip arrays, comparing whole-genome expression profiles for wild type mixed-stage embryos and for *tbx-2(bx59)* mutant animals (L. Clary and P. Okkema, personal communication). The initial analysis of the data was performed by The Core Genomics Facility, and resulted in approximately 1200 differentially expressed genes (out of approximately 20,000 genes tested) at a false discovery rate of 5%, approximately 1000 of which were up-regulated in the mutant (L. Clary and P. Okkema, personal communication).

These 1200 genes were then compared to existing data sets of pharyngeal expression. Approximately 240 genes were identified as preferentially expressed in the pharynx (Gaudet and Mango, 2002), including *myo-5*. As *myo-5* was down-regulated on the *tbx-2(bx59)* microarray (at an FDR of 5%), it is a potential candidate for tbx-2 acting on myo-5 as an activator in the pharynx. A reporter construct was built to assay expression in the mutant as compared to wild type.

Materials and Methods

Worm strains were maintained on nematode growth media (NGM) seeded with the OP50 *E. coli* strain (Brenner, 1974). The myo-5 promoter region was PCR amplified from N2 genomic DNA using primers PO939 and PO940. The *myo-5* promoter region was spliced to a

APPENDIX B (continued)

PCR amplified protein coding region for green fluorescent protein (GFP) by overlap extension during PCR (i.e. PCR Fusion) (Warrens et al., 1997) using primers PO941 and PO910. An electrophoresis gel was run confirming successful PCR fusion. Germline transformation via injection (Mello et al., 1991) was used to create worms with the myo-5::GFP extra-chromosomal array. The *rol-6(su1006)* co-injection marker was included for all transformants. Strains were maintained by passaging rolling worms. Several lines were recovered and the expression patterns were imaged using differential contrast and fluorescent imaging using a Zeiss Axioskop fluorescence microscope. Images were captured with AxioVision software.

Results

The expression pattern of *myo-5* is found in the pharyngeal muscles. Because GFP expression was cytoplasmic in these muscles, it could not be determined if expression was present in other pharyngeal cell types. Expression begins during the late embryo stage (Figure 31), and continues in the pharynx in adult worms (Figure 32). There are puncta of fluorescence along the isthmus which correspond to physical structures observable in DIC, and a higher concentration in the terminal bulb.

APPENDIX B (continued)

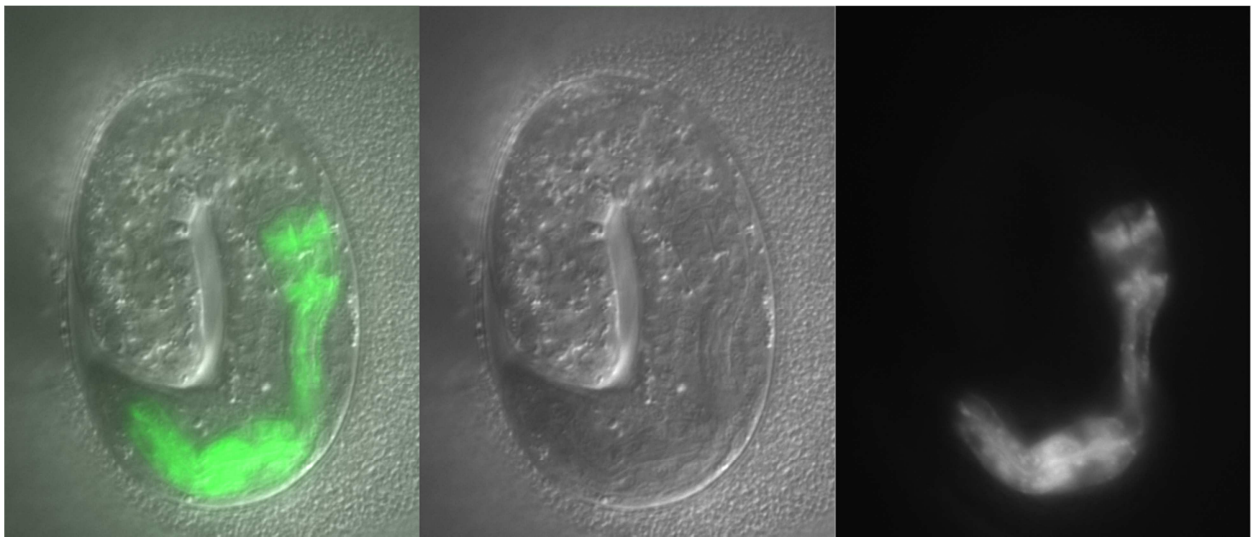

Figure 31: myo-5::GFP Expression in Late Embryo

Expression of myo-5::GFP in late embryo stages is detected in pharyngeal muscles. As the expression is cytoplasmic it could not be determined if expression was present in other pharyngeal cell types.

APPENDIX B (continued)

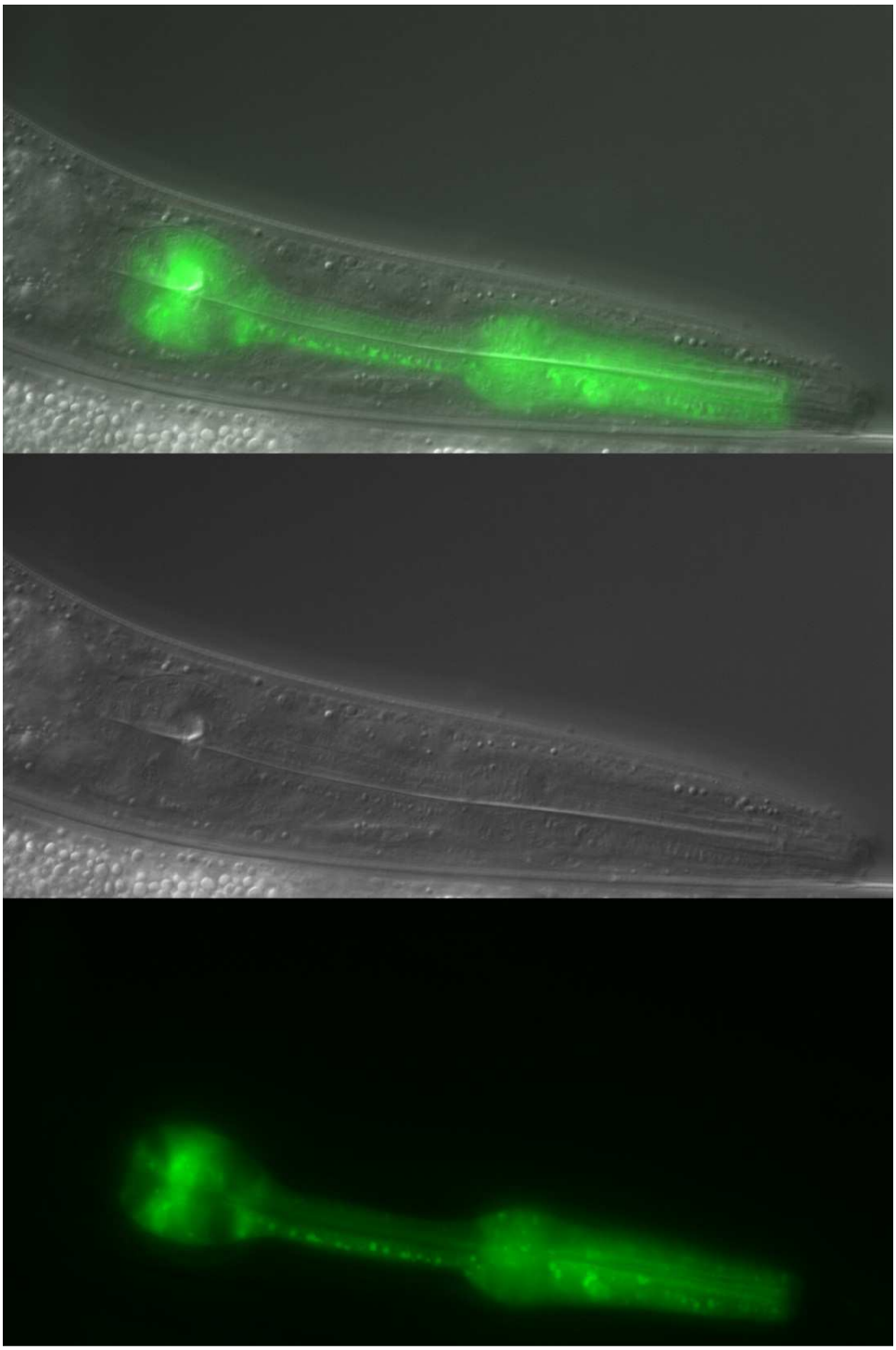

Figure 32: myo-5::GFP Expression in Adult

Expression of myo-5::GFP in adult worms is detected in pharyngeal muscles. As the expression is cytoplasmic it could not be determined if expression was present in other pharyngeal cell types.

APPENDIX B (continued)

Discussion

To test whether *myo-5* may be a target of *tbx-2*, one could cross this extra-chromosomal reporter into the mutant *tbx-2(bx59)* background, and shift to impermissive temperature. The expression patterns of the reporter could be compared, and differences noted. If *tbx-2* acts as a activator for *myo-5*, one would expect GFP expression to decrease or disappear. While this would be evidence that *myo-5* is downstream of *tbx-2*, direct interaction would need to be established by other means.

Appendix C

PROTEIN EXPRESSION ATTEMPT FOR TBX-2

Introduction

No sequence specific binding model currently exists for TBX-2 in *C. elegans* or any other species. In order to assay the DNA-binding motif for TBX-2, an attempt was made to express a maltose binding protein tagged full-length TBX-2 in a bacterial system. Initial expression attempts failed.

Materials and Methods

The plasmids pOK246.03, pOK246.04, and pOK246.05, and an empty vector pMa12C as a control were transformed into *E. coli* (BL21(DE3)) using the High Efficiency Bacterial Transformation protocol. BL21(DE3) competent cells were thawed on ice. 1uL of DNA was mixed with 100uL of competent cells on ice in 5mL tubes, and were then incubated for 40 minutes. Competent cells were heat-shocked for 20 seconds in a 42˚C water bath, and then placed on ice for 2 minutes. 100uL of 2xTY were then added to cells, and they were shaken at 37˚C, 225 rpm, for 1 hour. Bacteria were transferred to a 2xTY plate with ampicillin, and incubated overnight at 37˚C.

A single colony was used to inoculate 25mL of 2xTY medium containing 0.2% glucose and 100 µg/ml ampicillin. The culture was grown at 37˚C with shaking to mid-log phase (OD600 = 0.4 to 0.6). A 3 ml sample was transferred to a 15 mL snap cap tube and moved to 37˚C with shaking for an uninduced culture. A 1mL sample was removed as a pre-induction sample. Next, 63uL of 0.1M IPTG (isopropyl-1-thio-B-D-galactoside) (Invitrogen) was added to each culture in

order to induce protein production via the *lac* promoter. The culture was returned to 37˚C with shaking for 2 hours before a second 1mL sample was removed, and the process was repeated at 4 hours. Each collected 1mL sample was centrifuged at 14kRPM for 2 minutes, pelleted, and resuspended in 100uL of a 1xSDS sample buffer and stored on ice. Samples were then heated at 95˚C, centrifuged at 14kRPM for 2 minutes. For each sample, 15uL were run on a 10% SDS-PAGE gel at 200V, and then stained with Coomassie.

Results

No TBX-2 expression was detected at these experimental conditions (data not shown).

Discussion

TBX-2 expression failed at these experimental parameters, as well as despite several previous attempts at expression (P. Okkema, personal communication). Further research indicated that other successful attempts at expression of T-Box proteins were done using the Rosetta2 strain (Novagen) designed to enhance the expression of eukaryotic proteins that contain codons rarely used in *E. coli*. (Ghosh et al., 2001; Macindoe et al., 2009). A future attempt to express the binding domain of TBX-2 protein using the Rosetta2 strain was the first successful expression of *C. elegans* TBX-2 in bacteria (L. Clary, P. Okkema, personal communication).

Appendix D

BIOLOGICAL TESTING OF PREDICTED TBX-2 TARGETS

Introduction

The methods in Chapter 6 were used to create a pool of genes enriched for potential direct targets of TBX-2. At a loosened set of parameters (differential expression at 5% FDR with up regulation in the mutant; proximate input conservation of 3 or more; anti-correlation at $p < 0.01$), a total of 46 potential targets were predicted, 4 of which correlated at 20-minutes time lag. These four genes – *mxl-3*, *nex-1*, *F53F4.13*, and *Y60C6A.1* – were chosen for biological testing. The initial step of testing was to produce a reporter construct containing a high-copy number promoter-GFP fusion in an extrachromosomal which could be imaged in both wild-type and mutant backgrounds.

Materials and Methods

A plasmid containing 2.4kB of ceh-28 promoter in a pPD122.56 backbone was received from the Horvitz lab (Hirose et al., 2010). The plasmid was sequenced and renamed pOK288.04.

Shortened promoter regions of potential targets were PCR amplified from N2 genomic DNA using primers PO1325 and PO1326 (for *mxl-3*), and primers PO1327 and PO1328 (for *nex-1*), primers PO1329 and PO1330 (for *F53F4.13*), and primers PO1331 and PO1332 (for *Y60C6A.1*). These primers were designed to create an overlap with the pOK288.04 backbone for use with the In-Fusion cloning protocol. pOK288.04 was digested with SmaI (New England Biolabs) using NEB4 buffer at 25°C for 2 hours, and then with HindIII-HF (New England Biolabs)

using NEB4 buffer at 37°C for 16 hours. The digest was heat inactivated at 80°C for 20 minutes to deactivate the restriction enzymes.

Successfully PCR amplified promoter region inserts were PCR purified, and DNA concentrations were assayed using a Nanodrop spectrophotometer (Thermo Scientific). A 2:1 molar ratio of vector to insert was calculated. In each In-Fusion reaction, 2uL of 5X In-Fusion HD Enzyme Premix, approximately 180ng of linearized vector DNA, insert DNA and dH_2O were added to 10uL total volume. For the 523bp *F53F4.13* promoter region, 33.5ng of DNA was used. For the 2016bp promoter region of *mxl-3*, 93ng of DNA was used. For the 1876bp promoter region of *nex-1*, 84ng of DNA was used. The In-Fusions reactions were incubated at 50°C for 15 minutes, placed on ice, and then transformed into Stellar competent cells (Clontech) using the Stellar Competent Cell Protocol (Clontech).

Results

The PCR linearization of the pOK288.04 vector backbone was completed successfully. PCR amplification from N2 genomic DNA was successfully completed for *mxl-3*, *nex-1*, and *F53F4.13*. No clones were successfully created.

Appendix E

INITIAL PREDICTION OF A TBX-2 BINDING SITE (NEWPWM)

Introduction

No sequence specific binding model currently exists for TBX-2 in *C. elegans* or any other species. In order to predict the DNA-binding motif for TBX-2 in *C. elegans*, promoter regions from potential downstream targets (differentially expressed genes from a tbx-2(bx59) mutant microarray) were mined for patterns and a potential binding motif for TBX-2 in *C. elegans* was created.

Materials and Methods

A microarray experiment was performed using Affymetrix *C. elegans* GeneChip arrays, comparing whole-genome expression profiles for wild type mixed-stage embryos and for *tbx-2(bx59)* mutant animals (L. Clary and P. Okkema, personal communication). The initial analysis of the data was performed by The Core Genomics Facility, and resulted in approximately 1200 differentially expressed genes (out of approximately 20,000 genes tested) at a false discovery rate of 5%, approximately 1000 of which were up regulated in the mutant (L. Clary and P. Okkema, personal communication).

Using the Brachyury motif from JASPAR (Portales-Casamar et al., 2010), the promoter regions of differentially expressed genes were searched for potential binding sites within 500 base pairs of the start codon. A list of 257 binding site sequences (from the most statistically

significant differentially expressed up-regulated genes) was compiled and used to create a new sequence motif.

Results

The newly created motif (newPWM) differed from the Brachyury motif in several ways (Figure 33). The preferences at position 1 for cytosine (C) and position 9 for guanine (G) were almost completely eliminated. The strength of prediction at any individual location was decreased, but positions 3, 6 and 10 were decreased less than other positions. Overall, the sequence logo shows a similar motif.

Discussion

By creating a new motif based on a related T-Box motif – but drawn from only genes downstream of TBX-2 – it was hypothesized that key traits would be seen where TBX-2 binding differed from that of Brachyury. While this method is ultimately flawed, primarily due to being created from a pool of both direct and indirect targets, some interesting patterns can be seen.

By building a model from the results of another model based on an imperfect match, noise is introduced. Since the first model is compared against the genome with a less-than-100% threshold, imperfect matches will be found and incorporated into the result set. While these imperfect matches allow us to find a new yet related pattern, they also introduce random noise into the results at each position. We attempt to reduce this random noise by assembling the new model from sites that are more likely to be active: only the subset of matches found close to the start codon of differentially expressed genes. In an ideal scenario when a new set is built from 100% verified and biologically active targets, the only positions that differ are due to

APPENDIX E (continued)

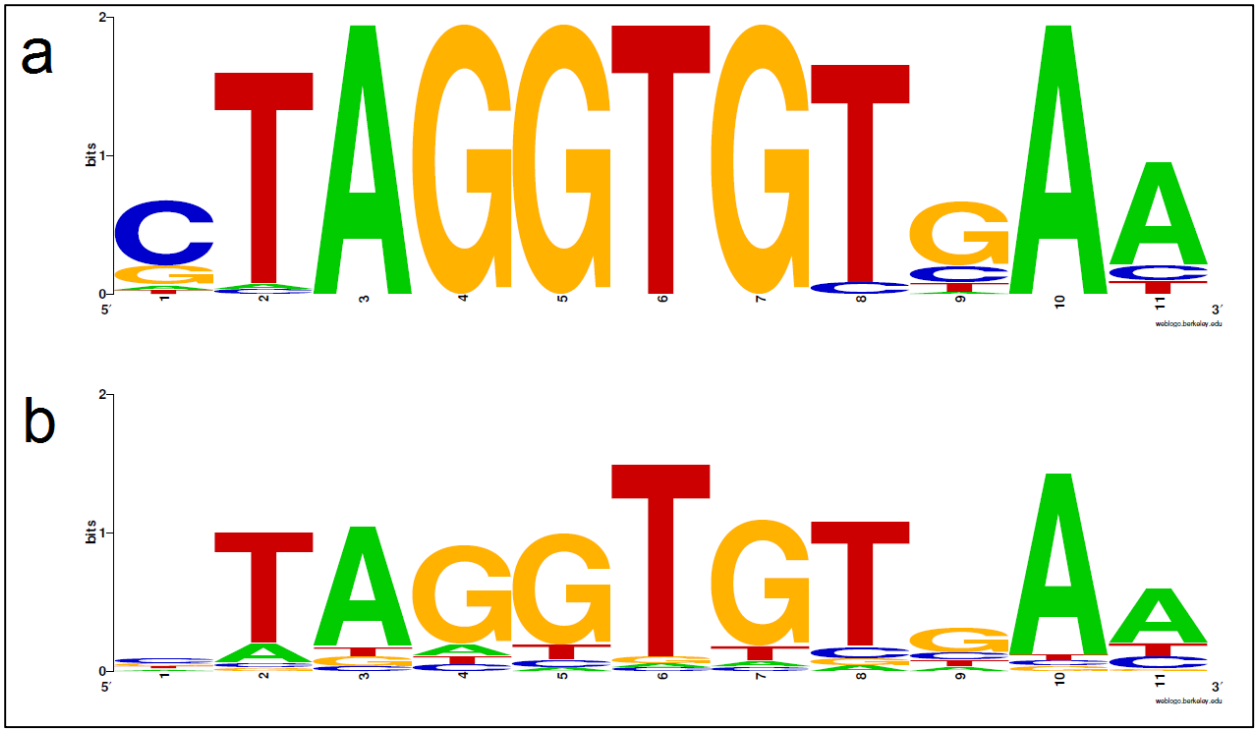

Figure 33: Brachyury and newPWM Sequence Logos

Sequence logo comparison between (a) the Brachyury motif (JASPAR) and (b) The newPWM motif.

functional differences between the sites, with no difference due to noise. However, as the set of genes differentially expressed on a microarray contains indirect targets as well as direct targets, it will be a source of noise.

Despite seeing an overall degradation of quality in the motif (visualized by smaller letter height across the board) some positions change less than others. These differences are likely unrelated to noise, and more likely to be related to binding motif in *C. elegans*. Based on these results, and an assumption of a constant introduction of noise at each position, positions 6 and 10 are disproportionately preserved, and seem to be present despite introduction of noise into the model. Similarly position three seems to have decreased less than positions 4 or 5. This implies these positions are more prevalent in the *C. elegans* genome or that they may be important for TBX-2 binding in *C. elegans*.

Appendix F

SERIAL ANALYSIS OF GENE EXPRESSION (SAGE) FOR TBX-2

Introduction

While genomics presents a static picture of what potential genes exist, the transcriptome tells us about the dynamic behavior and expression of those genes. While more recent methods of measuring gene expression rely on decreasing costs of sequencing (Martin and Wang, 2011), one of the first comprehensive methods was Serial Analysis of Gene Expression (SAGE) (Velculescu et al., 1995; 1997) Using methods of fluorescent cell sorting (FACS), pools of cells can be significantly enriched for specific organs, regions, or even individual cells (McKay et al., 2003). These pools are then subjected to SAGE, resulting in transcript-level expression information about these cell populations.

Here, the *C. elegans* SAGE data sets are used to determine in which tissues tbx-2 is enriched.

Materials and Methods

C. elegans SAGE data and mappings to WormBase release WS190 were obtained from the Genome BC C. elegans Gene Expression Consortium (http://elegans.bcgsc.bc.ca/). For the LongSage data set, the tag counts for pharynx cells were compared to FACS sorted AFD neurons, ASER neurons, ciliated neurons gut cells, hypodermal cells, muscle cells, pan-neural cells, pharyngeal marginal cells, whole N2 Embryos, and purified oocytes. The statistical

significance for each tag comparison was calculated (Kal et al., 1999), and multiple test correction was applied (Benjamini, 1995).

Results

The SAGE tags in pharyngeal cells were compared to the SAGE tags found in other cell subsets, whole embryos, and purified oocytes. Tags enriched in pharyngeal cells as compared to other categories were determined at a false discovery rate of 5% (Figure 34). Several hundred enriched tags were found in comparison to each category. The largest number of statistically significant SAGE tags was found in comparison with the subset derived from pharyngeal marginal cells, and the second largest number of enriched tags was found in oocytes.

Discussion

In this analysis, the SAGE results are useful for determining which transcripts are preferentially expressed in pharyngeal cells. The raw SAGE data is also useful for determining the relative numbers of all transcripts in pharyngeal cells. The large number of significant numbers of cells when compared to pharyngeal marginal cells is surprising, as one would expect a smaller number of transcripts to be significant between the pharynx and a proper subset of the pharynx. This analysis would be most useful when one were attempting to ask which transcripts are preferentially pharyngeal in contrast to another set of cells. For example, since TBX-2 is known to be expressed in neurons as well as the pharynx, one could use the tags enhanced in the pharynx in contrast with neurons in order eliminate neural targets, or to select for pharyngeal targets.

APPENDIX F (continued)

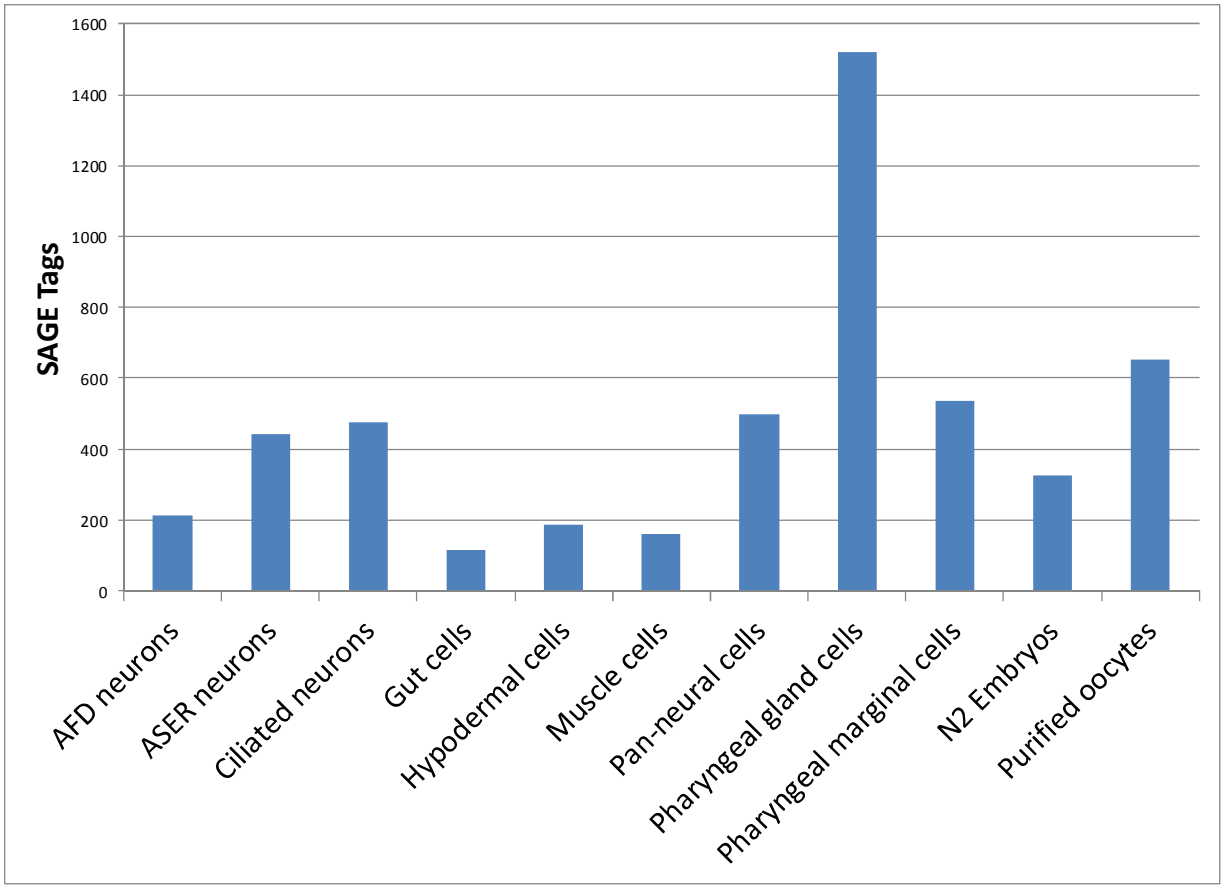

Figure 34: SAGE Tags Enriched in Pharyngeal Cells By Category (at 5% FDR)
Enriched SAGE tags for pharyngeal cells as compared to other group available in the LongSAGE *C. elegans* data set (at a FDR of 5%).

Appendix G

PREDICTING POTENTIAL UPSTREAM ACTIVATORS OF TBX-2 USING MULAN

Introduction

An enhancer for *tbx-2* was identified, dubbed CR2, which contains the regulatory elements required for pharyngeal expression (A. Milton and P. Okkema, personal communication) (Figure 35). This region, when attached to a basal promoter, is necessary and sufficient to drive pharyngeal expression of a reporter construct. CR2 is located upstream of tbx-2 (Figure 36), in the penultimate intron of the next upstream gene (Figure 37). CR2 encompasses a highly conserved region (Figure 38), and should contain regulatory motifs vital to tbx-2 expression in the pharynx.

In order to identify these motifs, the MULAN suite of bioinformatics tools (Loots and Ovcharenko, 2007) was used to identify transcription factor binding site motifs in conserved regions. Several potential upstream regulators of *tbx-2* were identified, including a FoxC1 binding site which matches the binding motif of a known upstream regulator of pharyngeal genes, *pha-4*.

Materials and Methods

A three-way genome alignment containing the extended region around CR2 was created using *C. briggsae*, *C. remanei*, and *C. elegans* genomic sequences with ClustalW2 (Thompson et al., 2002). These sequences were submitted to MULAN, aligned, and parameters of ECR length

APPENDIX G (continued)

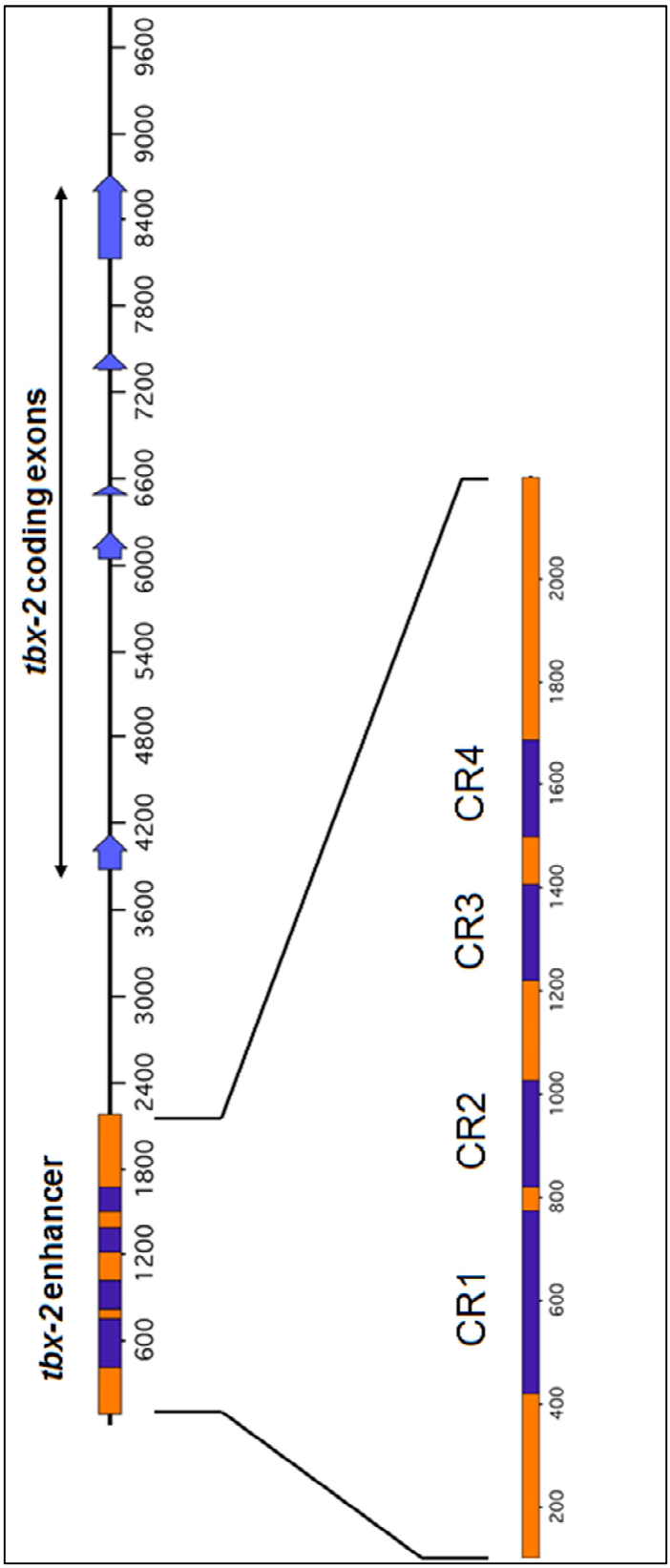

Figure 35: CR2 Schematic
A schematic of the region surrounding *C. elegans tbx-2* shows the upstream enhancer region containing conserved regions CR1 through CR4.

APPENDIX G (continued)

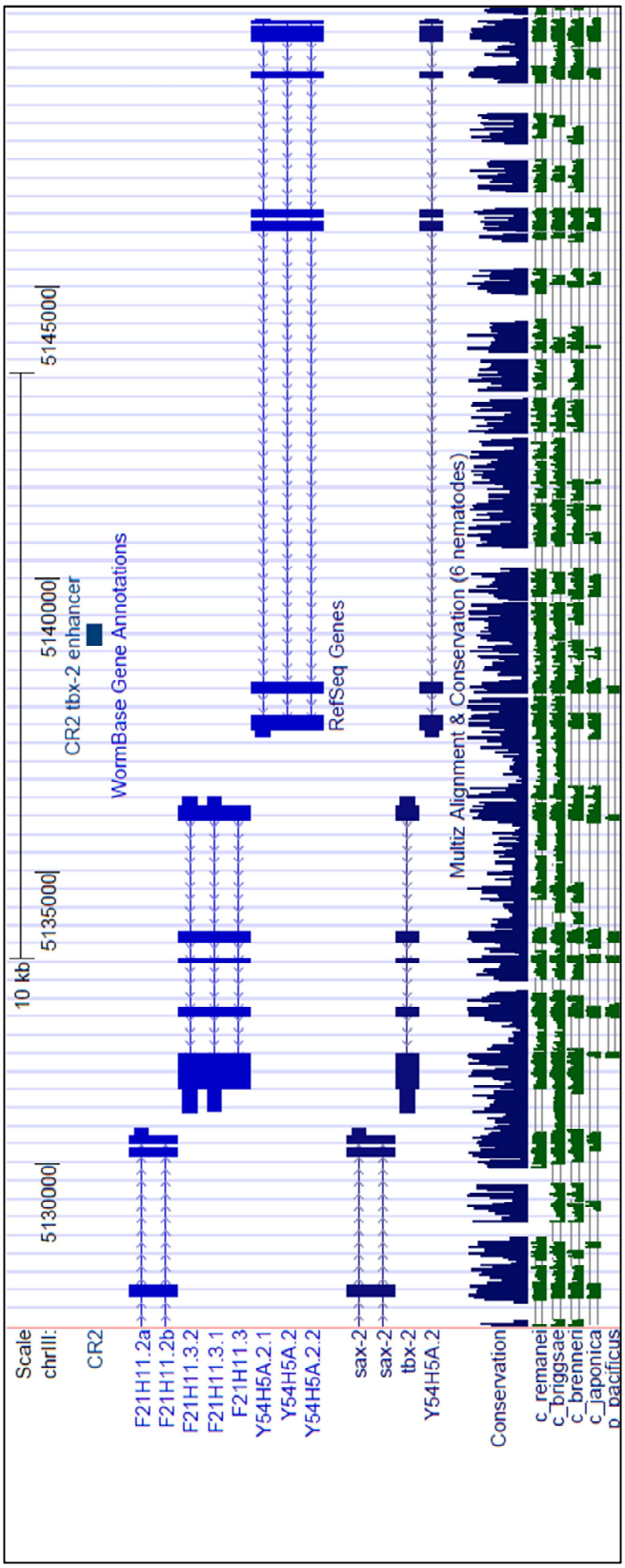

Figure 36: Genomic Context of the CR2 Enhancer

The genomic region around *C. Elegans tbx-2* on chromosome III is shown. Coding regions are indicated by dark blue horizontal bars and intronic regions by thin, hashed bars. Conservation is graphically indicated below, with the height of the graphic relating to phastCons conservation score (from 0 to 1). t*bx-2* is located lefot of center, and is on the negative strand so it is arranged right-to-left. The CR2 enhancer is marked by a dark blue rectangle in the CR2 track and is located in the penultimate intron of the next gene upstream.

APPENDIX G (continued)

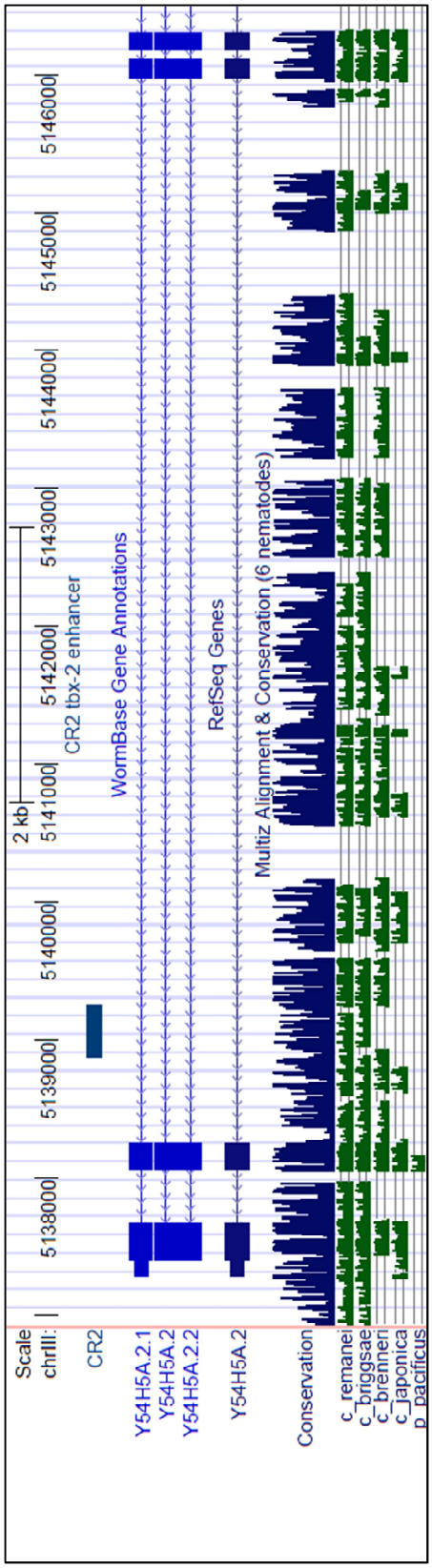

Figure 37: CR2 is Located in the Penultimate Intron of Upstream Gene
The genomic region around *C. Elegans tbx-2* enhancer CR2 on chromosome III is shown. Coding regions are indicated by dark blue horizontal bars and intronic regions by thin, hashed bars. Conservation is graphically indicated below, with the height of the graphic relating to phastCons conservation score (from 0 to 1). The CR2 enhancer is marked by a dark blue rectangle in the CR2 track.

APPENDIX G (continued)

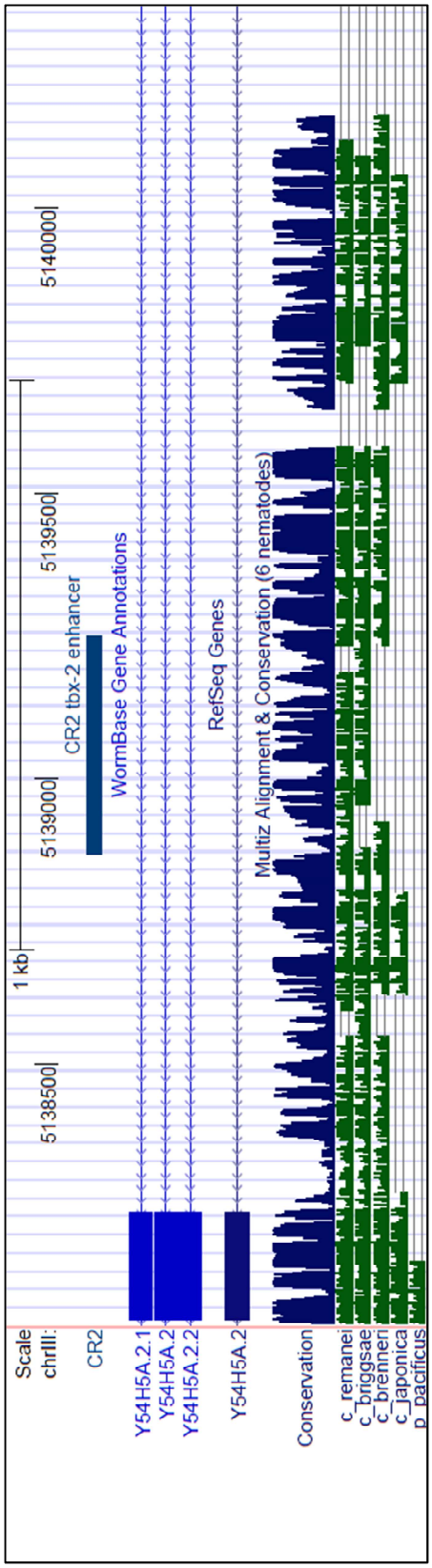

Figure 38: Conservation of the CR2 Region
The genomic region around *C. Elegans tbx-2* enhancer CR2 on chromosome III is shown. Coding regions are indicated by dark blue horizontal bars and intronic regions by thin, hashed bars. Conservation is graphically indicated below, with the height of the graphic relating to phastCons conservation score (from 0 to 1). The CR2 enhancer is marked by a dark blue rectangle in the CR2 track.

of 100 and threshold of 70% were chosen. The MULAN suite then searched the conserved region CR2 for matches to binding site motifs using TRANSFAC professional (version 10.2). Matches were visualized on a map of the aligned region (Figure 39) with CR2 marked in purple.

Results

Several potential motifs were identified, including a forkhead box motif FoxC1, and motifs for vertebrate NCX, MYB, E2F, as well as *C. elegans ces-2* (Figure 40). The forkhead box sequence matched, TATTTAC, is a match for the *pha-4* binding consensus, TRTTKRY or T(A/G)TT(T/G)(A/G)(T/C) known to be required in *C. elegans* for expression of pharyngeal genes. NCX is a transcription factor involved in patterning and regulation of early heart development (Linask et al., 2001). MYB, an acronym for myeloblastosis, is an oncogene, and while Myb is believed to have been lost from the *C. elegans* genome (Ganter and Lipsick, 1999), 19 genes encoding Myb-related DNA-binding domains are predicted in the *C. elegans* genome. E2F genes are a group of transcription factors in higher eukaryotes involved in the cell cycle regulation and synthesis of DNA, some of which act as activators and others as repressors. E2F transcription factors bind to the TTTCGCGC consensus binding site. *ces-2* encodes a basic region leucine zipper (bZIP) transcription factor, and is involved in controlling programmed cell death (Metzstein et al., 1996).

APPENDIX G (continued)

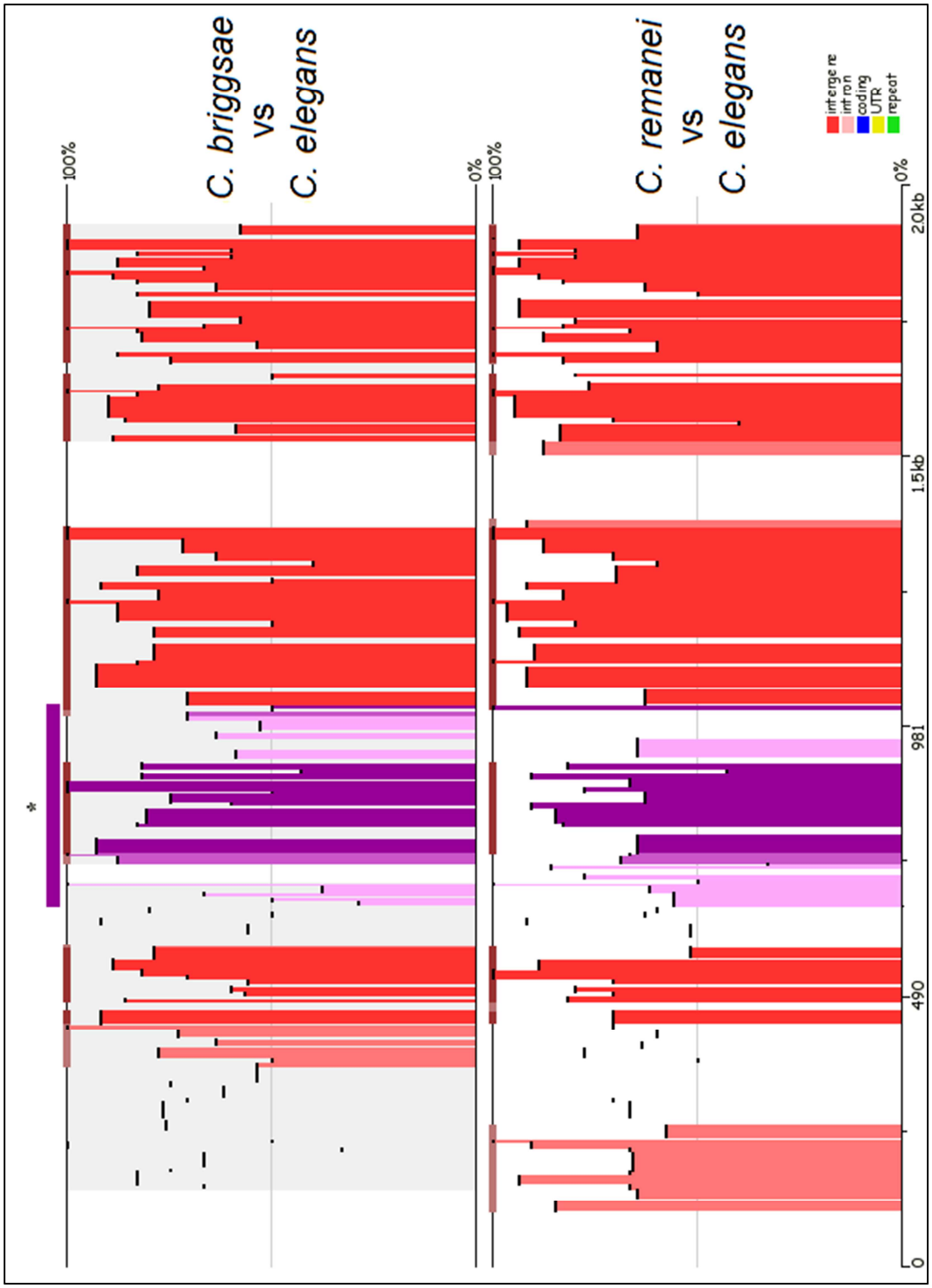

Figure 39: MULAN Alignment Results
Graphical output of the MULAN alignments of the *tbx-2* pharyngeal enhancer show several conserved regions. Conserved region 2 (CR2) is marked in purple. Bar height refers to conservation amount (calculated by moving window averages).

APPENDIX G (continued)

Figure 40: Conserved Transcription Factor Binding Sites in CR2

Conserved binding motifs found in the region of the *tbx-2* pharyngeal enhancer. Each colored rectangle marks a motif hit. (Note: multiple motifs may find hits on the same sequence.) Motif hits by class are labeled on the left in black. The conserved region 2 (CR2) is marked in purple.

APPENDIX G (continued)

Discussion

Both the FoxC1 motif and the *ces-2* motif seem to fit know characteristics of pharyngeal expression of *tbx-2*. The FoxC1 motif matches a sequence for *pha-4* binding, a gene known to be required for all pharyngeal expression. The *ces-2* motif is interesting, as the exact fate of the missing anterior pharyngeal cells in *tbx-2* mutants is not known. Although the typical markers for apoptosis were not seen in tbx-2 mutants (Smith and Mango, 2007), further analysis may be required to completely rule out programmed cell death, especially if the region of this motif is found to influence *tbx-2* expression.

Appendix H

PREDICTING POTENTIAL UPSTREAM ACTIVATORS OF TBX-2 USING CORRELATION

Introduction

As discussed above in Chapter 5, *tbx-2* expression begins to increase at the 100-cell stage, and peaks towards the end of the embryonic time course microarray data set (Gaudet and Mango, 2002). As shown above, the data set can be used to predict potential target genes which *lag* in time, but the predictive capabilities and statistical significance of the data set are limited. In contrast, the data set can be used for genes which *lead* in time with improved results due to the late peak of *tbx-2*.

Here, a method is presented where potential upstream activator transcription factors are predicted by finding gene expression profiles which correlate positively with *tbx-2* but which do so at various time-*leads*.

Materials and Methods

Cross-correlation was calculated (Baugh et al., 2005) using expression values at each time point for *tbx-2* and each gene in the embryonic time course microarray data set for 20-minutes lead, 40-minutes lead, and 60-minutes lead using the *ccf* function in Bioconductor (Kato and Tsunoda, 2001). Two probes in the microarray map to the genomic region for *tbx-2*, one in a coding region, and one in the 3'-UTR. Both had substantially similar expression patterns, but the coding region probe showed higher expression. The coding region probe was used for subsequent analysis. Plots were generated using the *gplot* function for the cross-correlation function, and for the expression levels for *tbx-2* and each gene tested. Correlation

coefficients were evaluated for statistical significance (Gentleman et al., 2004). The list of genes positively correlating at various lead times was limited to potential *C. elegans* transcription factors (Kato and Tsunoda, 2001).

Results

Of 20,000 *C. elegans* genes, 934 are identified as potential transcription factors (Reece-Hoyes et al., 2005). Of these 934 transcription factors, 97 correlate positively with *tbx-2* at various lead times ($p<0.01$) (**Error! Reference source not found.**). At a 60-minute lead time, the transcription factor *mec-3* is the only positive correlator ($r=0.805$, $p < 0.01$). At 40-minutes lead time, 26 transcription factors have strong positive correlation ($r > .75$, $p < 0.01$) including *fkh-2* and *pes-1*, both forkhead transcription factors. At 20-minutes lead time, 70 transcription factors have strong positive correlation ($r > .71$, $p < 0.01$) including *fkh-2*, *fkh-7*, and *pha-4*.

Discussion

This method presents a list of potential upstream activators of *tbx-2*. While the only two criteria applied are positive correlation at various lead times and presence on a list of *C. elegans* transcription factors, additional criteria could be used to improve these predictions. Each potential transcription factor could be assayed for expression in the pharynx, and a literature review could be conducted to determine whether the transcription factor can act as an activator. Some predicted candidates are unlikely to be upstream activators, despite being identified by these methods. For example, *gei-17* is a known participant in the SUMOylation pathway in *C. elegans*, but is not known to be a direct activator of transcription. Thus further review of this predicted list would be useful.

Gene	pcc (r)	p (<0.01)	Lead
mec-3	0.805	0.008	60
atg-4.1	0.840	0.002	40
C38D4.7	0.840	0.002	40
ceh-44	0.784	0.006	40
ceh-44	0.763	0.008	40
daf-3	0.757	0.009	40
die-1	0.766	0.008	40
dve-1	0.775	0.007	40
elt-1	0.766	0.008	40
fkh-2	0.788	0.006	40
grl-25	0.826	0.003	40
K02G10.1	0.753	0.010	40
K10B3.5	0.755	0.009	40
lin-40	0.777	0.007	40
lir-3	0.790	0.006	40
lsy-2	0.751	0.010	40
mml-1	0.756	0.009	40
nhr-201	0.827	0.003	40
nhr-202	0.827	0.003	40
nhr-43	0.794	0.005	40
pes-1	0.811	0.004	40
psa-1	0.773	0.007	40
R06F6.6	0.858	0.002	40
ref-2	0.764	0.008	40
scrt-1	0.804	0.004	40
syd-9	0.760	0.009	40
unc-62	0.764	0.008	40
ztf-16	0.822	0.003	40
ada-2	0.738	0.007	20
alr-1	0.765	0.005	20
atf-6	0.787	0.003	20
atg-4.1	0.724	0.009	20
B0250.4	0.787	0.003	20
C01B12.2	0.729	0.008	20
C02F5.12	0.811	0.002	20

Gene	pcc (r)	p (<0.01)	Lead
C07E3.6	0.783	0.004	20
C34F6.9	0.823	0.002	20
cdc-14	0.772	0.004	20
cdc-14	0.742	0.007	20
cdc-14	0.720	0.009	20
ceh-27	0.820	0.002	20
cey-1	0.736	0.008	20
cey-1	0.735	0.008	20
cey-1	0.723	0.009	20
cnd-1	0.790	0.003	20
dhhc-1	0.802	0.003	20
dmd-7	0.735	0.008	20
egl-27	0.732	0.008	20
elt-3	0.795	0.003	20
elt-3	0.753	0.006	20
eor-1	0.751	0.006	20
F16B12.6	0.756	0.006	20
F21A10.2	0.794	0.003	20
F21A10.2	0.768	0.005	20
F21A9.2	0.894	0.000	20
F23F12.9	0.833	0.001	20
F23F12.9	0.820	0.002	20
F57C9.4	0.747	0.006	20
fkh-2	0.818	0.002	20
fkh-2	0.766	0.005	20
fkh-7	0.814	0.002	20
gei-17	0.766	0.005	20
hmg-1.1	0.741	0.007	20
hmg-1.2	0.723	0.009	20
irx-1	0.806	0.002	20
K02G10.1	0.786	0.004	20
klf-3	0.825	0.002	20
let-526	0.736	0.008	20
lin-28	0.771	0.005	20
lin-28	0.756	0.006	20

Gene	pcc (r)	p (<0.01)	Lead
lin-32	0.753	0.006	20
lsy-2	0.736	0.008	20
mdl-1	0.910	0.000	20
mep-1	0.729	0.008	20
mls-2	0.829	0.002	20
mls-2	0.828	0.002	20
mls-2	0.826	0.002	20
nhr-1	0.741	0.007	20
nhr-205	0.888	0.000	20
nhr-232	0.720	0.009	20
nhr-28	0.860	0.001	20
nhr-28	0.757	0.006	20
nhr-33	0.806	0.002	20
nhr-35	0.852	0.001	20
nhr-42	0.792	0.003	20
nhr-46	0.784	0.004	20
nhr-46	0.774	0.004	20
nhr-46	0.773	0.004	20
nhr-47	0.792	0.003	20
nhr-53	0.849	0.001	20
nhr-57	0.814	0.002	20
nhr-64	0.725	0.009	20
nob-1	0.826	0.002	20
nob-1	0.777	0.004	20
odd-2	0.817	0.002	20
pax-3	0.917	0.000	20
pha-4	0.756	0.006	20
psa-1	0.823	0.002	20
set-16	0.743	0.007	20
sma-9	0.726	0.009	20
sptf-1	0.759	0.005	20
syd-9	0.806	0.002	20
syd-9	0.778	0.004	20
T04G9.1	0.721	0.009	20
T23G5.6	0.754	0.006	20

Gene	pcc (r)	p (<0.01)	Lead
tag-192	0.772	0.004	20
tag-192	0.735	0.008	20
tlp-1	0.763	0.005	20
unc-120	0.731	0.008	20
unc-130	0.774	0.004	20
unc-39	0.742	0.007	20
vab-15	0.851	0.001	20
vab-15	0.799	0.003	20
Y37E11B.1	0.767	0.005	20
ztf-13	0.767	0.005	20
ztf-16	0.767	0.005	20

Figure 41: Potential Upstream Activators of *tbx-2*

Potential upstream activators of tbx-2 were selected by high positive correlation ($p < 0.01$) with expression patterns of *C. elegans* transcription factors at 20-, 40-, and 60-minute lead times.

APPENDIX H (continued)

In addition, any predictions by this method which also arise under the MULAN method in Appendix G would have a stronger foundation. The different methods used should not share the same false positive predictions, so the combination of methods might provide a more reliable prediction. For example, the MULAN method in Appendix G predicts a forkhead transcription factor binding site in the conserved CR2 region. While *pha-4*, a pharyngeal specifying transcription factor is one candidate, other forkhead transcription factors may interact at this site. Several forkhead transcription factors are predicted as potential upstream regulators of *tbx-2*, and should be considered as potentially upstream regulators interacting at this conserved binding motif.

Appendix I

GENE ONTOLOGY TERMS ENRICHED FOR DIFFERENTIALLY EXPRESSED GENES

Gene to GO Cell Compartment test for over-representation

GOCCID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0030288	0.003	21.491	0	3	5	outer membrane-bounded periplasmic space
GO:0042597	0.003	21.491	0	3	5	periplasmic space
GO:0030313	0.005	14.325	0	3	6	cell envelope
GO:0044462	0.005	14.325	0	3	6	external encapsulating structure part
GO:0030312	0.008	10.743	0	3	7	external encapsulating structure
GO:0044428	0.018	2.234	5	11	82	nuclear part
GO:0005576	0.026	1.852	9	15	132	extracellular region

Gene to GO Molecular Function test for over-representation

GOMFID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0004857	0	3.425	5	14	74	enzyme inhibitor activity
GO:0030414	0	3.465	4	13	68	peptidase inhibitor activity
GO:0004866	0.001	3.312	4	12	65	endopeptidase inhibitor activity
GO:0004497	0.001	2.785	6	15	94	monooxygenase activity
GO:0016705	0.001	4.101	3	9	41	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
GO:0005200	0.001	43.374	0	3	4	structural constituent of cytoskeleton
GO:0016491	0.002	1.676	30	47	467	oxidoreductase activity
GO:0051213	0.002	11.584	1	4	9	dioxygenase activity
GO:0016717	0.002	21.684	0	3	5	oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water
GO:0016715	0.003	9.652	1	4	10	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced ascorbate as one donor, and incorporation of one atom of oxygen
GO:0004867	0.004	3.037	4	10	58	serine-type endopeptidase inhibitor activity
GO:0004504	0.004	Inf	0	2	2	peptidylglycine monooxygenase activity

APPENDIX I (continued)

GO:0016702	0.005	14.454	0	3	6	oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen
GO:0043167	0.007	1.337	91	113	1406	ion binding
GO:0043169	0.007	1.337	91	113	1406	cation binding
GO:0016701	0.008	10.839	0	3	7	oxidoreductase activity, acting on single donors with incorporation of molecular oxygen
GO:0046872	0.009	1.323	89	109	1364	metal ion binding
GO:0004602	0.012	8.67	1	3	8	glutathione peroxidase activity
GO:0004365	0.012	28.853	0	2	3	glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) activity
GO:0004652	0.012	28.853	0	2	3	polynucleotide adenylyltransferase activity
GO:0008943	0.012	28.853	0	2	3	glyceraldehyde-3-phosphate dehydrogenase activity
GO:0005515	0.012	1.269	131	153	2015	protein binding
GO:0005102	0.014	2.767	3	8	50	receptor binding
GO:0005179	0.021	3.62	2	5	25	hormone activity
GO:0005544	0.023	14.425	0	2	4	calcium-dependent phospholipid binding
GO:0016290	0.023	14.425	0	2	4	palmitoyl-CoA hydrolase activity
GO:0016813	0.023	14.425	0	2	4	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amidines
GO:0005529	0.025	2.672	3	7	45	sugar binding
GO:0003887	0.031	5.416	1	3	11	DNA-directed DNA polymerase activity
GO:0070566	0.037	9.615	0	2	5	adenylyltransferase activity
GO:0008158	0.037	3.015	2	5	29	hedgehog receptor activity
GO:0034061	0.039	4.814	1	3	12	DNA polymerase activity

Gene to GO Biological Process test for over-representation

GOBPID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0010033	0.001	5.281	2	7	27	response to organic substance
GO:0010038	0.002	12.018	1	4	9	response to metal ion
GO:0006631	0.002	3.58	3	9	47	fatty acid metabolic process
GO:0010035	0.002	10.013	1	4	10	response to inorganic substance
GO:0010564	0.003	4.398	2	7	31	regulation of cell cycle process
GO:0070887	0.003	5.02	2	6	24	cellular response to chemical stimulus
GO:0001676	0.004	Inf	0	2	2	long-chain fatty acid metabolic process
GO:0008156	0.004	Inf	0	2	2	negative regulation of DNA replication
GO:0042759	0.004	Inf	0	2	2	long-chain fatty acid biosynthetic process

APPENDIX I (continued)

GO:0006633	0.004	5.784	1	5	18	fatty acid biosynthetic process
GO:0006801	0.004	14.993	0	3	6	superoxide metabolic process
GO:0051053	0.004	14.993	0	3	6	negative regulation of DNA metabolic process
GO:0032787	0.006	2.852	4	10	63	monocarboxylic acid metabolic process
GO:0042221	0.006	2.09	9	18	149	response to chemical stimulus
GO:0006800	0.007	11.243	0	3	7	oxygen and reactive oxygen species metabolic process
GO:0006082	0.008	1.892	12	21	190	organic acid metabolic process
GO:0019752	0.008	1.892	12	21	190	carboxylic acid metabolic process
GO:0043436	0.008	1.892	12	21	190	oxoacid metabolic process
GO:0006260	0.009	3.402	2	7	38	DNA replication
GO:0016053	0.01	3.295	2	7	39	organic acid biosynthetic process
GO:0046394	0.01	3.295	2	7	39	carboxylic acid biosynthetic process
GO:0010948	0.011	8.993	1	3	8	negative regulation of cell cycle process
GO:0009266	0.011	3.194	3	7	40	response to temperature stimulus
GO:0009074	0.011	29.93	0	2	3	aromatic amino acid family catabolic process
GO:0014055	0.011	29.93	0	2	3	acetylcholine secretion
GO:0031123	0.011	29.93	0	2	3	RNA 3'-end processing
GO:0043631	0.011	29.93	0	2	3	RNA polyadenylation
GO:0046688	0.011	29.93	0	2	3	response to copper ion
GO:0042180	0.012	1.815	12	21	197	cellular ketone metabolic process
GO:0015870	0.016	7.493	1	3	9	acetylcholine transport
GO:0055066	0.016	7.493	1	3	9	di-, tri-valent inorganic cation homeostasis
GO:0009408	0.018	3.755	2	5	25	response to heat
GO:0007218	0.021	6.422	1	3	10	neuropeptide signaling pathway
GO:0051052	0.021	6.422	1	3	10	regulation of DNA metabolic process
GO:0006275	0.022	14.963	0	2	4	regulation of DNA replication
GO:0019439	0.022	14.963	0	2	4	aromatic compound catabolic process
GO:0043055	0.022	14.963	0	2	4	maintenance of dauer
GO:0043405	0.022	14.963	0	2	4	regulation of MAP kinase activity
GO:0046686	0.022	14.963	0	2	4	response to cadmium ion
GO:0006950	0.026	1.66	13	21	213	response to stress
GO:0008286	0.028	5.618	1	3	11	insulin receptor signaling pathway
GO:0009072	0.028	5.618	1	3	11	aromatic amino acid family metabolic process
GO:0032868	0.028	5.618	1	3	11	response to insulin stimulus
GO:0032869	0.028	5.618	1	3	11	cellular response to insulin stimulus
GO:0043434	0.028	5.618	1	3	11	response to peptide hormone stimulus
GO:0007269	0.028	4	1	4	19	neurotransmitter secretion

APPENDIX I (continued)

GO:0016054	0.028	3.264	2	5	28	organic acid catabolic process
GO:0046395	0.028	3.264	2	5	28	carboxylic acid catabolic process
GO:0003001	0.033	3.749	1	4	20	generation of a signal involved in cell-cell signaling
GO:0009063	0.033	3.749	1	4	20	cellular amino acid catabolic process
GO:0006636	0.035	9.974	0	2	5	unsaturated fatty acid biosynthetic process
GO:0033261	0.035	9.974	0	2	5	regulation of S phase
GO:0034453	0.035	9.974	0	2	5	microtubule anchoring
GO:0051247	0.035	9.974	0	2	5	positive regulation of protein metabolic process
GO:0051313	0.035	9.974	0	2	5	attachment of spindle microtubules to chromosome
GO:0051320	0.035	9.974	0	2	5	S phase
GO:0055074	0.035	9.974	0	2	5	calcium ion homeostasis
GO:0009725	0.036	4.993	1	3	12	response to hormone stimulus
GO:0032870	0.036	4.993	1	3	12	cellular response to hormone stimulus
GO:0007067	0.037	3.002	2	5	30	mitosis
GO:0051726	0.039	2.392	3	7	51	regulation of cell cycle
GO:0040028	0.041	1.73	9	15	146	regulation of vulval development
GO:0048580	0.041	1.73	9	15	146	regulation of post-embryonic development
GO:0000280	0.042	2.886	2	5	31	nuclear division
GO:0009719	0.044	4.493	1	3	13	response to endogenous stimulus
GO:0032507	0.044	4.493	1	3	13	maintenance of protein location in cell
GO:0040020	0.044	4.493	1	3	13	regulation of meiosis
GO:0051651	0.044	4.493	1	3	13	maintenance of location in cell
GO:0060341	0.044	4.493	1	3	13	regulation of cellular localization
GO:0048285	0.047	2.779	2	5	32	organelle fission
GO:0033559	0.05	7.479	0	2	6	unsaturated fatty acid metabolic process
GO:0046626	0.05	7.479	0	2	6	regulation of insulin receptor signaling pathway
GO:0055065	0.05	7.479	0	2	6	metal ion homeostasis

APPENDIX I

Appendix J

GENE ONTOLOGY TERMS ENRICHED BY SUBCLUSTER

49 Gene Subcluster Gene to GO Cellular Compartment test for over-representation

GOCCID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0005634	0.019	5.263	1	4	872	nucleus
GO:0043229	0.021	4.562	2	5	1393	intracellular organelle
GO:0043226	0.021	4.558	2	5	1394	organelle
GO:0005875	0.037	30.524	0	1	29	microtubule associated complex
GO:0044424	0.043	3.667	2	5	1659	intracellular part
GO:0043231	0.048	3.84	1	4	1146	intracellular membrane-bounded organelle
GO:0043227	0.048	3.836	1	4	1147	membrane-bounded organelle

49 Gene Subcluster Gene to GO Molecular Function test for over-representation

GOMFID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0000166	0.007	3.77	3	8	1163	nucleotide binding
GO:0005524	0.014	3.738	2	6	791	ATP binding
GO:0032559	0.014	3.738	2	6	791	adenyl ribonucleotide binding
GO:0030554	0.019	3.458	2	6	847	adenyl nucleotide binding
GO:0001883	0.019	3.453	2	6	848	purine nucleoside binding
GO:0001882	0.019	3.435	2	6	852	nucleoside binding
GO:0032553	0.029	3.094	2	6	933	ribonucleotide binding
GO:0032555	0.029	3.094	2	6	933	purine ribonucleotide binding
GO:0017076	0.04	2.851	3	6	1001	purine nucleotide binding
GO:0004713	0.045	4.232	1	3	306	protein tyrosine kinase activity

49 Gene Subcluster Gene to GO Biological Process test for over-representation

GOBPID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0007049	0.037	4.601	1	3	307	cell cycle
GO:0043687	0.042	3.533	1	4	550	post-translational protein modification
GO:0032501	0.044	2.653	9	13	3625	multicellular organismal process
GO:0007017	0.044	6.582	0	2	138	microtubule-based process
GO:0007275	0.048	2.506	8	12	3251	multicellular organismal development
GO:0009790	0.049	2.432	6	10	2487	embryonic development

57 Gene Subcluster Gene to GO Cellular Compartment test for over-representation

APPENDIX I (continued)

GOCCID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0030288	0.012	113.353	0	1	5	outer membrane-bounded periplasmic space
GO:0042597	0.012	113.353	0	1	5	periplasmic space
GO:0030313	0.014	90.671	0	1	6	cell envelope
GO:0044462	0.014	90.671	0	1	6	external encapsulating structure part
GO:0030312	0.016	75.549	0	1	7	external encapsulating structure

57 Gene Subcluster Gene to GO Molecular Function test for over-representation

GOMFID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0005088	0.001	46.562	0	2	20	Ras guanyl-nucleotide exchange factor activity
GO:0005089	0.001	46.562	0	2	20	Rho guanyl-nucleotide exchange factor activity
GO:0030695	0.002	15.567	0	3	88	GTPase regulator activity
GO:0060589	0.002	15.205	0	3	90	nucleoside-triphosphatase regulator activity
GO:0005021	0.003	Inf	0	1	1	vascular endothelial growth factor receptor activity
GO:0005085	0.004	24.595	0	2	36	guanyl-nucleotide exchange factor activity
GO:0004713	0.008	6.04	1	4	306	protein tyrosine kinase activity
GO:0004652	0.008	198.333	0	1	3	polynucleotide adenylyltransferase activity
GO:0016772	0.011	4.492	1	5	531	transferase activity, transferring phosphorus-containing groups
GO:0030234	0.013	7.051	0	3	188	enzyme regulator activity
GO:0004714	0.013	99.139	0	1	5	transmembrane receptor protein tyrosine kinase activity
GO:0070566	0.013	99.139	0	1	5	adenylyltransferase activity
GO:0005083	0.014	12.239	0	2	70	small GTPase regulator activity
GO:0004674	0.015	4.853	1	4	376	protein serine/threonine kinase activity
GO:0004672	0.019	4.567	1	4	398	protein kinase activity
GO:0016773	0.027	4.062	1	4	444	phosphotransferase activity, alcohol group as acceptor
GO:0016301	0.029	3.938	1	4	457	kinase activity
GO:0019199	0.031	36.015	0	1	12	transmembrane receptor protein kinase activity

57 Gene Subcluster Gene to GO Biological Process test for over-representation

GOBPID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0051056	0.001	15.616	0	3	76	regulation of small GTPase mediated signal transduction

APPENDIX I (continued)

GO:0035023	0.002	38.284	0	2	21	regulation of Rho protein signal transduction
GO:0007266	0.002	34.629	0	2	23	Rho protein signal transduction
GO:0009966	0.005	9.684	0	3	120	regulation of signal transduction
GO:0010646	0.007	8.769	0	3	132	regulation of cell communication
GO:0031123	0.009	173.595	0	1	3	RNA 3'-end processing
GO:0043631	0.009	173.595	0	1	3	RNA polyadenylation
GO:0007242	0.011	5.309	1	4	297	intracellular signaling cascade
GO:0007264	0.012	7.177	0	3	160	small GTPase mediated signal transduction
GO:0030241	0.012	115.714	0	1	4	muscle thick filament assembly
GO:0031033	0.012	115.714	0	1	4	myosin filament assembly or disassembly
GO:0031034	0.012	115.714	0	1	4	myosin filament assembly
GO:0007155	0.012	13.16	0	2	57	cell adhesion
GO:0022610	0.012	13.16	0	2	57	biological adhesion
GO:0046578	0.016	11.663	0	2	64	regulation of Ras protein signal transduction
GO:0007265	0.02	10.029	0	2	74	Ras protein signal transduction
GO:0014866	0.021	57.833	0	1	7	skeletal myofibril assembly
GO:0006468	0.028	3.933	1	4	394	protein amino acid phosphorylation
GO:0007160	0.03	38.54	0	1	10	cell-matrix adhesion
GO:0031589	0.03	38.54	0	1	10	cell-substrate adhesion
GO:0040020	0.038	28.893	0	1	13	regulation of meiosis
GO:0051302	0.038	28.893	0	1	13	regulation of cell division
GO:0006464	0.039	3.063	2	5	644	protein modification process
GO:0007156	0.041	26.667	0	1	14	homophilic cell adhesion
GO:0043412	0.044	2.951	2	5	666	biopolymer modification
GO:0010927	0.044	24.759	0	1	15	cellular component assembly involved in morphogenesis
GO:0030239	0.044	24.759	0	1	15	myofibril assembly
GO:0031032	0.044	24.759	0	1	15	actomyosin structure organization
GO:0035058	0.044	24.759	0	1	15	sensory cilium assembly
GO:0051146	0.044	24.759	0	1	15	striated muscle cell differentiation
GO:0055002	0.044	24.759	0	1	15	striated muscle cell development
GO:0042048	0.047	23.105	0	1	16	olfactory behavior
GO:0055001	0.047	23.105	0	1	16	muscle cell development
GO:0016310	0.048	3.271	1	4	468	phosphorylation
GO:0051445	0.05	21.658	0	1	17	regulation of meiotic cell cycle

147 Gene Subcluster Gene to GO Cellular Compartment test for over-representation

GOCCID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0016021	0	19.261	34	47	5496	integral to membrane

APPENDIX I (continued)

GO:0031224	0	19.2	34	47	5501	intrinsic to membrane
GO:0044425	0	18.231	35	47	5582	membrane part
GO:0016020	0	16.543	36	47	5729	membrane

147 Gene Subcluster Gene to GO Molecular Function test for over-representation

GOMFID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0016491	0.001	4.516	2	9	467	oxidoreductase activity
GO:0008131	0.005	Inf	0	1	1	amine oxidase activity
GO:0004784	0.011	192.486	0	1	2	superoxide dismutase activity
GO:0016721	0.011	192.486	0	1	2	oxidoreductase activity, acting on superoxide radicals as acceptor
GO:0048038	0.011	192.486	0	1	2	quinone binding
GO:0004497	0.013	6.624	0	3	94	monooxygenase activity
GO:0005242	0.016	96.23	0	1	3	inward rectifier potassium channel activity
GO:0016641	0.026	48.101	0	1	5	oxidoreductase activity, acting on the CH-NH2 group of donors, oxygen as acceptor
GO:0005261	0.03	4.799	1	3	128	cation channel activity
GO:0020037	0.03	4.799	1	3	128	heme binding
GO:0046906	0.031	4.722	1	3	130	tetrapyrrole binding
GO:0016638	0.042	27.475	0	1	8	oxidoreductase activity, acting on the CH-NH2 group of donors
GO:0046873	0.046	4.012	1	3	152	metal ion transmembrane transporter activity
GO:0005267	0.046	6.226	0	2	65	potassium channel activity
GO:0008234	0.048	6.128	0	2	66	cysteine-type peptidase activity

147 Gene Subcluster Gene to GO Biological Process test for over-representation

GOBPID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0006801	0.024	50.207	0	1	6	superoxide metabolic process
GO:0030001	0.025	5.214	1	3	155	metal ion transport
GO:0015672	0.026	5.145	1	3	157	monovalent inorganic cation transport
GO:0006800	0.028	41.833	0	1	7	oxygen and reactive oxygen species metabolic process
GO:0006813	0.031	7.813	0	2	68	potassium ion transport
GO:0055114	0.032	4.707	1	3	171	oxidation reduction
GO:0006614	0.032	35.852	0	1	8	SRP-dependent cotranslational protein targeting to membrane
GO:0045047	0.032	35.852	0	1	8	protein targeting to ER
GO:0006810	0.036	2.413	4	8	962	transport
GO:0006613	0.036	31.366	0	1	9	cotranslational protein targeting to membrane

APPENDIX I (continued)

326 Gene Subcluster Gene to GO Cellular Compartment test for over-representation

GOCCID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0000323	0.005	25.78	0	2	7	lytic vacuole
GO:0005764	0.005	25.78	0	2	7	lysosome
GO:0031224	0.01	1.722	85	97	5501	intrinsic to membrane
GO:0016021	0.017	1.637	85	96	5496	integral to membrane
GO:0044425	0.019	1.634	87	97	5582	membrane part
GO:0005773	0.019	10.732	0	2	14	vacuole
GO:0016507	0.031	63.941	0	1	2	fatty acid beta-oxidation multienzyme complex
GO:0016020	0.033	1.566	89	98	5729	membrane
GO:0005874	0.041	6.772	0	2	21	microtubule

326 Gene Subcluster Gene to GO Molecular Function test for over-representation

GOMFID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0046872	0	2.486	18	34	1364	metal ion binding
GO:0043167	0	2.392	18	34	1406	ion binding
GO:0043169	0	2.392	18	34	1406	cation binding
GO:0046914	0	2.335	15	28	1129	transition metal ion binding
GO:0016491	0.001	2.815	6	15	467	oxidoreductase activity
GO:0004497	0.001	5.47	1	6	94	monooxygenase activity
GO:0008237	0.002	4.449	2	7	134	metallopeptidase activity
GO:0005102	0.004	6.861	1	4	50	receptor binding
GO:0005179	0.004	10.676	0	3	25	hormone activity
GO:0008270	0.004	2.055	12	22	948	zinc ion binding
GO:0004222	0.006	4.668	1	5	90	metalloendopeptidase activity
GO:0003707	0.015	2.689	3	8	247	steroid hormone receptor activity
GO:0016705	0.016	6.167	1	3	41	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
GO:0004879	0.017	2.621	3	8	253	ligand-dependent nuclear receptor activity
GO:0005506	0.017	3.117	2	6	159	iron ion binding
GO:0020037	0.025	3.208	2	5	128	heme binding
GO:0003857	0.026	76.815	0	1	2	3-hydroxyacyl-CoA dehydrogenase activity
GO:0004300	0.026	76.815	0	1	2	enoyl-CoA hydratase activity
GO:0046906	0.027	3.156	2	5	130	tetrapyrrole binding
GO:0003865	0.038	38.402	0	1	3	3-oxo-5-alpha-steroid 4-dehydrogenase activity
GO:0016229	0.038	38.402	0	1	3	steroid dehydrogenase activity

APPENDIX I (continued)

GO:0033765	0.038	38.402	0	1	3	steroid dehydrogenase activity, acting on the CH-CH group of donors
GO:0004867	0.039	4.25	1	3	58	serine-type endopeptidase inhibitor activity

326 Gene Subcluster Gene to GO Biological Process test for over-representation

GOBPID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0006631	0.002	8.803	1	4	47	fatty acid metabolic process
GO:0055114	0.003	4.134	2	7	171	oxidation reduction
GO:0044255	0.003	4.612	1	6	131	cellular lipid metabolic process
GO:0006629	0.003	3.914	2	7	180	lipid metabolic process
GO:0006082	0.005	3.695	2	7	190	organic acid metabolic process
GO:0019752	0.005	3.695	2	7	190	carboxylic acid metabolic process
GO:0043436	0.005	3.695	2	7	190	oxoacid metabolic process
GO:0032787	0.005	6.401	1	4	63	monocarboxylic acid metabolic process
GO:0042180	0.006	3.556	2	7	197	cellular ketone metabolic process
GO:0006633	0.016	11.569	0	2	18	fatty acid biosynthetic process
GO:0001676	0.022	91.57	0	1	2	long-chain fatty acid metabolic process
GO:0042759	0.022	91.57	0	1	2	long-chain fatty acid biosynthetic process
GO:0046627	0.022	91.57	0	1	2	negative regulation of insulin receptor signaling pathway
GO:0009408	0.03	8.04	0	2	25	response to heat
GO:0016054	0.037	7.109	0	2	28	organic acid catabolic process
GO:0046395	0.037	7.109	0	2	28	carboxylic acid catabolic process
GO:0019509	0.043	30.515	0	1	4	methionine salvage
GO:0043055	0.043	30.515	0	1	4	maintenance of dauer
GO:0043102	0.043	30.515	0	1	4	amino acid salvage
GO:0046686	0.043	30.515	0	1	4	response to cadmium ion
GO:0048585	0.043	30.515	0	1	4	negative regulation of response to stimulus

48 Gene Subcluster Gene to GO Cellular Compartment test for over-representation

GOCCID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0016585	0.002	Inf	0	1	1	chromatin remodeling complex
GO:0044428	0.017	11.222	0	2	82	nuclear part
GO:0005681	0.017	71.343	0	1	7	spliceosomal complex
GO:0005576	0.041	6.861	0	2	132	extracellular region

48 Gene Subcluster Gene to GO Molecular Function test for over-representation

GOMFID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0015662	0.001	47.645	0	2	15	ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism

APPENDIX I (continued)

GO:0042302	0.001	44.236	0	2	16	structural constituent of cuticle
GO:0003682	0.003	30.939	0	2	22	chromatin binding
GO:0031490	0.003	Inf	0	1	1	chromatin DNA binding
GO:0031491	0.003	Inf	0	1	1	nucleosome binding
GO:0042625	0.005	22.075	0	2	30	ATPase activity, coupled to transmembrane movement of ions
GO:0017111	0.006	5.013	1	5	344	nucleoside-triphosphatase activity
GO:0016462	0.007	4.804	1	5	358	pyrophosphatase activity
GO:0016818	0.007	4.79	1	5	359	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
GO:0016817	0.008	4.665	1	5	368	hydrolase activity, acting on acid anhydrides
GO:0042623	0.008	8.181	0	3	120	ATPase activity, coupled
GO:0004066	0.01	148.625	0	1	3	asparagine synthase (glutamine-hydrolyzing) activity
GO:0004365	0.01	148.625	0	1	3	glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) activity
GO:0008943	0.01	148.625	0	1	3	glyceraldehyde-3-phosphate dehydrogenase activity
GO:0005200	0.014	99.069	0	1	4	structural constituent of cytoskeleton
GO:0005388	0.014	99.069	0	1	4	calcium-transporting ATPase activity
GO:0005544	0.014	99.069	0	1	4	calcium-dependent phospholipid binding
GO:0015085	0.014	99.069	0	1	4	calcium ion transmembrane transporter activity
GO:0016887	0.017	6.101	1	3	159	ATPase activity
GO:0042626	0.018	10.612	0	2	60	ATPase activity, coupled to transmembrane movement of substances
GO:0043492	0.018	10.612	0	2	60	ATPase activity, coupled to movement of substances
GO:0005509	0.019	5.87	1	3	165	calcium ion binding
GO:0004012	0.021	59.425	0	1	6	phospholipid-translocating ATPase activity
GO:0005548	0.021	59.425	0	1	6	phospholipid transporter activity
GO:0015247	0.021	59.425	0	1	6	aminophospholipid transporter activity
GO:0015399	0.021	9.763	0	2	65	primary active transmembrane transporter activity
GO:0015405	0.021	9.763	0	2	65	P-P-bond-hydrolysis-driven transmembrane transporter activity
GO:0016820	0.024	9.175	0	2	69	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances

APPENDIX I (continued)

GO:0016620	0.024	49.514	0	1	7	oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor
GO:0016884	0.028	42.435	0	1	8	carbon-nitrogen ligase activity, with glutamine as amido-N-donor
GO:0016879	0.029	8.187	0	2	77	ligase activity, forming carbon-nitrogen bonds
GO:0004386	0.031	7.972	0	2	79	helicase activity
GO:0016903	0.031	37.125	0	1	9	oxidoreductase activity, acting on the aldehyde or oxo group of donors
GO:0043566	0.034	32.995	0	1	10	structure-specific DNA binding
GO:0015082	0.038	29.692	0	1	11	di-, tri-valent inorganic cation transmembrane transporter activity

48 Gene Subcluster Gene to GO Biological Process test for over-representation

GOBPID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0006986	0	93.346	0	2	8	response to unfolded protein
GO:0034620	0	93.346	0	2	8	cellular response to unfolded protein
GO:0051789	0	93.346	0	2	8	response to protein stimulus
GO:0009607	0.001	55.977	0	2	12	response to biotic stimulus
GO:0070887	0.004	25.402	0	2	24	cellular response to chemical stimulus
GO:0034514	0.004	Inf	0	1	1	mitochondrial unfolded protein response
GO:0043044	0.004	Inf	0	1	1	ATP-dependent chromatin remodeling
GO:0022414	0.004	3.278	4	10	1066	reproductive process
GO:0010033	0.005	22.345	0	2	27	response to organic substance
GO:0006338	0.008	269.852	0	1	2	chromatin remodeling
GO:0045104	0.008	269.852	0	1	2	intermediate filament cytoskeleton organization
GO:0045109	0.008	269.852	0	1	2	intermediate filament organization
GO:0006996	0.009	4.469	1	5	343	organelle organization
GO:0040035	0.009	3.468	2	7	646	hermaphrodite genitalia development
GO:0048806	0.01	3.438	2	7	651	genitalia development
GO:0006528	0.011	134.907	0	1	3	asparagine metabolic process
GO:0006529	0.011	134.907	0	1	3	asparagine biosynthetic process
GO:0045103	0.011	134.907	0	1	3	intermediate filament-based process
GO:0007548	0.015	3.122	3	7	710	sex differentiation
GO:0006874	0.015	89.926	0	1	4	cellular calcium ion homeostasis
GO:0009067	0.015	89.926	0	1	4	aspartate family amino acid biosynthetic process
GO:0003006	0.019	2.99	3	7	738	reproductive developmental process
GO:0000302	0.019	67.435	0	1	5	response to reactive oxygen species
GO:0006875	0.019	67.435	0	1	5	cellular metal ion homeostasis

APPENDIX I (continued)

GO:0055074	0.019	67.435	0	1	5	calcium ion homeostasis
GO:0006754	0.021	9.757	0	2	59	ATP biosynthetic process
GO:0046034	0.022	9.424	0	2	61	ATP metabolic process
GO:0015914	0.023	53.941	0	1	6	phospholipid transport
GO:0030104	0.023	53.941	0	1	6	water homeostasis
GO:0055065	0.023	53.941	0	1	6	metal ion homeostasis
GO:0009142	0.023	9.265	0	2	62	nucleoside triphosphate biosynthetic process
GO:0009145	0.023	9.265	0	2	62	purine nucleoside triphosphate biosynthetic process
GO:0009201	0.023	9.265	0	2	62	ribonucleoside triphosphate biosynthetic process
GO:0009206	0.023	9.265	0	2	62	purine ribonucleoside triphosphate biosynthetic process
GO:0009144	0.025	8.964	0	2	64	purine nucleoside triphosphate metabolic process
GO:0009199	0.025	8.964	0	2	64	ribonucleoside triphosphate metabolic process
GO:0009205	0.025	8.964	0	2	64	purine ribonucleoside triphosphate metabolic process
GO:0009141	0.025	8.821	0	2	65	nucleoside triphosphate metabolic process
GO:0006984	0.026	44.944	0	1	7	ER-nuclear signaling pathway
GO:0030005	0.026	44.944	0	1	7	cellular di-, tri-valent inorganic cation homeostasis
GO:0030968	0.026	44.944	0	1	7	endoplasmic reticulum unfolded protein response
GO:0034976	0.026	44.944	0	1	7	response to endoplasmic reticulum stress
GO:0009152	0.028	8.416	0	2	68	purine ribonucleotide biosynthetic process
GO:0009260	0.029	8.166	0	2	70	ribonucleotide biosynthetic process
GO:0009150	0.03	8.047	0	2	71	purine ribonucleotide metabolic process
GO:0006270	0.03	38.519	0	1	8	DNA replication initiation
GO:0033554	0.031	7.931	0	2	72	cellular response to stress
GO:0048856	0.031	2.334	5	10	1411	anatomical structure development
GO:0009259	0.031	7.818	0	2	73	ribonucleotide metabolic process
GO:0000003	0.032	2.215	7	12	1855	reproduction
GO:0009066	0.034	33.699	0	1	9	aspartate family amino acid metabolic process
GO:0055066	0.034	33.699	0	1	9	di-, tri-valent inorganic cation homeostasis
GO:0006164	0.035	7.397	0	2	77	purine nucleotide biosynthetic process
GO:0050896	0.036	2.757	3	6	662	response to stimulus

APPENDIX I (continued)

GO:0030003	0.038	29.951	0	1	10	cellular cation homeostasis
GO:0006163	0.038	7.019	0	2	81	purine nucleotide metabolic process
GO:0006325	0.041	6.759	0	2	84	chromatin organization
GO:0040002	0.041	6.677	0	2	85	collagen and cuticulin-based cuticle development
GO:0002119	0.043	2.138	7	11	1704	nematode larval development
GO:0002164	0.043	2.136	7	11	1705	larval development
GO:0007592	0.043	6.518	0	2	87	protein-based cuticle development
GO:0042335	0.043	6.518	0	2	87	cuticle development
GO:0051716	0.043	6.518	0	2	87	cellular response to stimulus
GO:0009791	0.046	2.107	7	11	1723	post-embryonic development
GO:0006950	0.047	4.044	1	3	213	response to stress

145 Gene Subcluster Gene to GO Cellular Compartment test for over-representation

GOCCID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0016021	0.008	2.452	40	48	5496	integral to membrane
GO:0031224	0.009	2.444	40	48	5501	intrinsic to membrane
GO:0044425	0.013	2.32	40	48	5582	membrane part
GO:0016020	0.028	2.105	42	48	5729	membrane

145 Gene Subcluster Gene to GO Molecular Function test for over-representation

GOMFID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0005199	0.008	Inf	0	1	1	structural constituent of cell wall
GO:0003951	0.015	134.075	0	1	2	NAD+ kinase activity
GO:0004084	0.015	134.075	0	1	2	branched-chain-amino-acid transaminase activity
GO:0008134	0.016	11.351	0	2	26	transcription factor binding
GO:0005544	0.03	44.679	0	1	4	calcium-dependent phospholipid binding
GO:0016813	0.03	44.679	0	1	4	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amidines
GO:0004356	0.037	33.505	0	1	5	glutamate-ammonia ligase activity
GO:0016211	0.044	26.8	0	1	6	ammonia ligase activity
GO:0016880	0.044	26.8	0	1	6	acid-ammonia (or amide) ligase activity
GO:0005529	0.045	6.318	0	2	45	sugar binding

145 Gene Subcluster Gene to GO Biological Process test for over-representation

GOBPID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0030497	0.007	Inf	0	1	1	fatty acid elongation
GO:0014055	0.02	75.667	0	1	3	acetylcholine secretion
GO:0043193	0.02	75.667	0	1	3	positive regulation of gene-specific transcription

APPENDIX I (continued)

GO:0009081	0.027	50.438	0	1	4	branched chain family amino acid metabolic process
GO:0060142	0.027	50.438	0	1	4	regulation of syncytium formation by plasma membrane fusion
GO:0016053	0.028	8.314	0	2	39	organic acid biosynthetic process
GO:0046394	0.028	8.314	0	2	39	carboxylic acid biosynthetic process
GO:0006542	0.033	37.823	0	1	5	glutamine biosynthetic process
GO:0006636	0.033	37.823	0	1	5	unsaturated fatty acid biosynthetic process
GO:0051605	0.033	37.823	0	1	5	protein maturation by peptide bond cleavage
GO:0006520	0.037	4.363	1	3	110	cellular amino acid metabolic process
GO:0044106	0.037	4.363	1	3	110	cellular amine metabolic process
GO:0006082	0.038	3.384	1	4	190	organic acid metabolic process
GO:0019752	0.038	3.384	1	4	190	carboxylic acid metabolic process
GO:0043436	0.038	3.384	1	4	190	oxoacid metabolic process
GO:0000768	0.04	30.254	0	1	6	syncytium formation by plasma membrane fusion
GO:0006949	0.04	30.254	0	1	6	syncytium formation
GO:0032583	0.04	30.254	0	1	6	regulation of gene-specific transcription
GO:0033559	0.04	30.254	0	1	6	unsaturated fatty acid metabolic process
GO:0046626	0.04	30.254	0	1	6	regulation of insulin receptor signaling pathway
GO:0042180	0.042	3.258	1	4	197	cellular ketone metabolic process

186 Gene Subcluster Gene to GO Cellular Compartment test for over-representation

GOCCID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0044428	0	8.8	1	6	82	nuclear part
GO:0005634	0.001	2.56	8	18	872	nucleus
GO:0005622	0.001	2.109	21	33	2148	intracellular
GO:0043231	0.002	2.3	11	21	1146	intracellular membrane-bounded organelle
GO:0043227	0.002	2.298	11	21	1147	membrane-bounded organelle
GO:0044446	0.004	3.017	4	10	387	intracellular organelle part
GO:0044422	0.004	3.008	4	10	388	organelle part
GO:0043229	0.004	2.069	13	23	1393	intracellular organelle
GO:0043226	0.004	2.067	13	23	1394	organelle
GO:0005654	0.004	10.067	0	3	35	nucleoplasm
GO:0044451	0.004	10.067	0	3	35	nucleoplasm part
GO:0044424	0.005	1.998	16	26	1659	intracellular part
GO:0031975	0.007	5.776	1	4	79	envelope
GO:0031981	0.007	8.471	0	3	41	nuclear lumen

APPENDIX I (continued)

GO:0016591	0.008	17.694	0	2	14	DNA-directed RNA polymerase II, holoenzyme
GO:0005663	0.01	Inf	0	1	1	DNA replication factor C complex
GO:0030141	0.01	Inf	0	1	1	secretory granule
GO:0031225	0.01	Inf	0	1	1	anchored to membrane
GO:0046658	0.01	Inf	0	1	1	anchored to plasma membrane
GO:0043233	0.013	6.697	0	3	51	organelle lumen
GO:0070013	0.013	6.697	0	3	51	intracellular organelle lumen
GO:0005635	0.014	12.482	0	2	19	nuclear envelope
GO:0031974	0.016	6.061	1	3	56	membrane-enclosed lumen
GO:0005657	0.019	104.863	0	1	2	replication fork
GO:0044427	0.023	5.261	1	3	64	chromosomal part
GO:0005667	0.024	9.219	0	2	25	transcription factor complex
GO:0016023	0.025	8.833	0	2	26	cytoplasmic membrane-bounded vesicle
GO:0031410	0.027	8.479	0	2	27	cytoplasmic vesicle
GO:0031988	0.027	8.479	0	2	27	membrane-bounded vesicle
GO:0000152	0.028	52.425	0	1	3	nuclear ubiquitin ligase complex
GO:0005672	0.028	52.425	0	1	3	transcription factor TFIIA complex
GO:0005680	0.028	52.425	0	1	3	anaphase-promoting complex
GO:0017053	0.028	52.425	0	1	3	transcriptional repressor complex
GO:0070176	0.028	52.425	0	1	3	DRM complex
GO:0031982	0.029	8.152	0	2	28	vesicle
GO:0031967	0.032	4.579	1	3	73	organelle envelope
GO:0043234	0.045	2.276	3	7	343	protein complex
GO:0005694	0.046	3.951	1	3	84	chromosome
GO:0030288	0.047	26.205	0	1	5	outer membrane-bounded periplasmic space
GO:0042597	0.047	26.205	0	1	5	periplasmic space

186 Gene Subcluster Gene to GO Molecular Function test for over-representation

GOMFID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0016702	0	87.333	0	3	6	oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen
GO:0016701	0	65.491	0	3	7	oxidoreductase activity, acting on single donors with incorporation of molecular oxygen
GO:0051213	0	43.648	0	3	9	dioxygenase activity
GO:0003887	0	32.727	0	3	11	DNA-directed DNA polymerase activity
GO:0034061	0	29.086	0	3	12	DNA polymerase activity

APPENDIX I (continued)

GO:0003702	0.008	8.154	0	3	35	RNA polymerase II transcription factor activity
GO:0003689	0.012	Inf	0	1	1	DNA clamp loader activity
GO:0004070	0.012	Inf	0	1	1	aspartate carbamoyltransferase activity
GO:0004086	0.012	Inf	0	1	1	carbamoyl-phosphate synthase activity
GO:0004370	0.012	Inf	0	1	1	glycerol kinase activity
GO:0004411	0.012	Inf	0	1	1	homogentisate 1,2-dioxygenase activity
GO:0004708	0.012	Inf	0	1	1	MAP kinase kinase activity
GO:0004712	0.012	Inf	0	1	1	protein serine/threonine/tyrosine kinase activity
GO:0004833	0.012	Inf	0	1	1	tryptophan 2,3-dioxygenase activity
GO:0016743	0.012	Inf	0	1	1	carboxyl- or carbamoyltransferase activity
GO:0018738	0.012	Inf	0	1	1	S-formylglutathione hydrolase activity
GO:0033170	0.012	Inf	0	1	1	protein-DNA loading ATPase activity
GO:0005515	0.017	1.663	24	33	2015	protein binding
GO:0016779	0.023	5.312	1	3	52	nucleotidyltransferase activity
GO:0003868	0.023	85.253	0	1	2	4-hydroxyphenylpyruvate dioxygenase activity
GO:0008191	0.023	85.253	0	1	2	metalloendopeptidase inhibitor activity
GO:0010576	0.023	85.253	0	1	2	metalloenzyme regulator activity
GO:0017017	0.023	85.253	0	1	2	MAP kinase tyrosine/serine/threonine phosphatase activity
GO:0033549	0.023	85.253	0	1	2	MAP kinase phosphatase activity
GO:0048551	0.023	85.253	0	1	2	metalloenzyme inhibitor activity
GO:0003682	0.027	8.606	0	2	22	chromatin binding
GO:0004190	0.029	8.195	0	2	23	aspartic-type endopeptidase activity
GO:0070001	0.029	8.195	0	2	23	aspartic-type peptidase activity
GO:0017056	0.035	42.62	0	1	3	structural constituent of nuclear pore
GO:0004866	0.04	4.191	1	3	65	endopeptidase inhibitor activity
GO:0005509	0.044	2.736	2	5	165	calcium ion binding
GO:0030414	0.045	3.995	1	3	68	peptidase inhibitor activity
GO:0005544	0.046	28.41	0	1	4	calcium-dependent phospholipid binding

186 Gene Subcluster Gene to GO Biological Process test for over-representation

GOBPID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0048513	0	2.378	13	26	966	organ development
GO:0009072	0	28.156	0	3	11	aromatic amino acid family metabolic process
GO:0009074	0.001	148.763	0	2	3	aromatic amino acid family catabolic process
GO:0046688	0.001	148.763	0	2	3	response to copper ion
GO:0006996	0.001	3.154	5	13	343	organelle organization

APPENDIX I (continued)

GO:0048731	0.001	2.237	14	26	1017	system development
GO:0019439	0.001	74.371	0	2	4	aromatic compound catabolic process
GO:0043405	0.001	74.371	0	2	4	regulation of MAP kinase activity
GO:0042325	0.001	16.076	0	3	17	regulation of phosphorylation
GO:0006260	0.002	8.894	1	4	38	DNA replication
GO:0019220	0.002	15.002	0	3	18	regulation of phosphate metabolic process
GO:0051174	0.002	15.002	0	3	18	regulation of phosphorus metabolic process
GO:0034453	0.002	49.574	0	2	5	microtubule anchoring
GO:0051313	0.002	49.574	0	2	5	attachment of spindle microtubules to chromosome
GO:0032502	0.002	1.813	46	60	3372	developmental process
GO:0048856	0.003	1.928	19	31	1411	anatomical structure development
GO:0040027	0.003	4.545	1	6	107	negative regulation of vulval development
GO:0022414	0.003	2.004	14	25	1066	reproductive process
GO:0048519	0.004	2.766	4	11	323	negative regulation of biological process
GO:0034641	0.004	5.065	1	5	80	cellular nitrogen compound metabolic process
GO:0006541	0.005	24.777	0	2	8	glutamine metabolic process
GO:0022402	0.005	2.825	4	10	286	cell cycle process
GO:0000165	0.006	21.234	0	2	9	MAPKKK cascade
GO:0010038	0.006	21.234	0	2	9	response to metal ion
GO:0000278	0.006	3.915	2	6	123	mitotic cell cycle
GO:0051093	0.007	3.881	2	6	124	negative regulation of developmental process
GO:0040011	0.007	1.85	17	27	1243	locomotion
GO:0007067	0.007	8.321	0	3	30	mitosis
GO:0010035	0.008	18.577	0	2	10	response to inorganic substance
GO:0000280	0.008	8.022	0	3	31	nuclear division
GO:0007049	0.008	2.618	4	10	307	cell cycle
GO:0048285	0.009	7.745	0	3	32	organelle fission
GO:0006725	0.009	5.288	1	4	61	cellular aromatic compound metabolic process
GO:0006519	0.009	3.573	2	6	134	cellular amino acid and derivative metabolic process
GO:0007275	0.01	1.639	44	56	3251	multicellular organismal development
GO:0040007	0.01	1.648	32	43	2337	growth
GO:0000087	0.012	7.016	0	3	35	M phase of mitotic cell cycle
GO:0006367	0.013	13.505	0	2	13	transcription initiation from RNA polymerase II promoter
GO:0007243	0.013	13.505	0	2	13	protein kinase cascade

APPENDIX I (continued)

GO:0032507	0.013	13.505	0	2	13	maintenance of protein location in cell
GO:0043549	0.013	13.505	0	2	13	regulation of kinase activity
GO:0045859	0.013	13.505	0	2	13	regulation of protein kinase activity
GO:0051338	0.013	13.505	0	2	13	regulation of transferase activity
GO:0051651	0.013	13.505	0	2	13	maintenance of location in cell
GO:0000188	0.014	Inf	0	1	1	inactivation of MAPK activity
GO:0001934	0.014	Inf	0	1	1	positive regulation of protein amino acid phosphorylation
GO:0006569	0.014	Inf	0	1	1	tryptophan catabolic process
GO:0006570	0.014	Inf	0	1	1	tyrosine metabolic process
GO:0007132	0.014	Inf	0	1	1	meiotic metaphase I
GO:0010562	0.014	Inf	0	1	1	positive regulation of phosphorus metabolic process
GO:0019441	0.014	Inf	0	1	1	tryptophan catabolic process to kynurenine
GO:0034514	0.014	Inf	0	1	1	mitochondrial unfolded protein response
GO:0042327	0.014	Inf	0	1	1	positive regulation of phosphorylation
GO:0042436	0.014	Inf	0	1	1	indole derivative catabolic process
GO:0045937	0.014	Inf	0	1	1	positive regulation of phosphate metabolic process
GO:0046218	0.014	Inf	0	1	1	indolalkylamine catabolic process
GO:0048144	0.014	Inf	0	1	1	fibroblast proliferation
GO:0048145	0.014	Inf	0	1	1	regulation of fibroblast proliferation
GO:0048147	0.014	Inf	0	1	1	negative regulation of fibroblast proliferation
GO:0051315	0.014	Inf	0	1	1	attachment of spindle microtubules to kinetochore during mitosis
GO:0051323	0.014	Inf	0	1	1	metaphase
GO:0040028	0.014	3.261	2	6	146	regulation of vulval development
GO:0048580	0.014	3.261	2	6	146	regulation of post-embryonic development
GO:0022403	0.014	2.707	3	8	235	cell cycle phase
GO:0000070	0.015	12.378	0	2	14	mitotic sister chromatid segregation
GO:0009064	0.015	12.378	0	2	14	glutamine family amino acid metabolic process
GO:0042221	0.015	3.191	2	6	149	response to chemical stimulus
GO:0007517	0.015	6.233	1	3	39	muscle organ development
GO:0040025	0.016	2.86	3	7	194	vulval development
GO:0006520	0.016	3.602	1	5	110	cellular amino acid metabolic process
GO:0044106	0.016	3.602	1	5	110	cellular amine metabolic process
GO:0007010	0.018	3.081	2	6	154	cytoskeleton organization
GO:0009308	0.018	3.081	2	6	154	amine metabolic process

APPENDIX I (continued)

GO:0006576	0.019	10.607	0	2	16	biogenic amine metabolic process
GO:0045185	0.019	10.607	0	2	16	maintenance of protein location
GO:0051235	0.019	10.607	0	2	16	maintenance of location
GO:0051656	0.019	4.178	1	4	76	establishment of organelle localization
GO:0003006	0.02	1.868	10	17	738	reproductive developmental process
GO:0051640	0.021	4.064	1	4	78	organelle localization
GO:0000819	0.022	9.898	0	2	17	sister chromatid segregation
GO:0002119	0.025	1.584	23	32	1704	nematode larval development
GO:0002164	0.025	1.582	23	32	1705	larval development
GO:0051276	0.026	3.172	2	5	124	chromosome organization
GO:0040035	0.026	1.864	9	15	646	hermaphrodite genitalia development
GO:0044271	0.027	4.98	1	3	48	nitrogen compound biosynthetic process
GO:0000187	0.027	73.622	0	1	2	activation of MAPK activity
GO:0001820	0.027	73.622	0	1	2	serotonin secretion
GO:0006469	0.027	73.622	0	1	2	negative regulation of protein kinase activity
GO:0006558	0.027	73.622	0	1	2	L-phenylalanine metabolic process
GO:0006559	0.027	73.622	0	1	2	L-phenylalanine catabolic process
GO:0006568	0.027	73.622	0	1	2	tryptophan metabolic process
GO:0006586	0.027	73.622	0	1	2	indolalkylamine metabolic process
GO:0006595	0.027	73.622	0	1	2	polyamine metabolic process
GO:0006596	0.027	73.622	0	1	2	polyamine biosynthetic process
GO:0006837	0.027	73.622	0	1	2	serotonin transport
GO:0007140	0.027	73.622	0	1	2	male meiosis
GO:0008584	0.027	73.622	0	1	2	male gonad development
GO:0008608	0.027	73.622	0	1	2	attachment of spindle microtubules to kinetochore
GO:0015844	0.027	73.622	0	1	2	monoamine transport
GO:0016574	0.027	73.622	0	1	2	histone ubiquitination
GO:0031134	0.027	73.622	0	1	2	sister chromatid biorientation
GO:0031401	0.027	73.622	0	1	2	positive regulation of protein modification process
GO:0032270	0.027	73.622	0	1	2	positive regulation of cellular protein metabolic process
GO:0033522	0.027	73.622	0	1	2	histone H2A ubiquitination
GO:0033673	0.027	73.622	0	1	2	negative regulation of kinase activity
GO:0042430	0.027	73.622	0	1	2	indole and derivative metabolic process
GO:0042434	0.027	73.622	0	1	2	indole derivative metabolic process
GO:0043086	0.027	73.622	0	1	2	negative regulation of catalytic activity
GO:0043406	0.027	73.622	0	1	2	positive regulation of MAP kinase activity
GO:0043407	0.027	73.622	0	1	2	negative regulation of MAP kinase activity
GO:0044092	0.027	73.622	0	1	2	negative regulation of molecular function

APPENDIX I (continued)

GO:0051348	0.027	73.622	0	1	2	negative regulation of transferase activity
GO:0048806	0.028	1.847	9	15	651	genitalia development
GO:0007548	0.028	1.812	10	16	710	sex differentiation
GO:0009791	0.029	1.561	23	32	1723	post-embryonic development
GO:0009063	0.029	8.245	0	2	20	cellular amino acid catabolic process
GO:0006352	0.032	7.81	0	2	21	transcription initiation
GO:0042692	0.032	7.81	0	2	21	muscle cell differentiation
GO:0000279	0.034	2.408	3	7	228	M phase
GO:0016043	0.036	1.861	8	13	555	cellular component organization
GO:0001703	0.038	4.305	1	3	55	gastrulation with mouth forming first
GO:0006575	0.038	7.064	0	2	23	cellular amino acid derivative metabolic process
GO:0009310	0.038	7.064	0	2	23	amine catabolic process
GO:0042127	0.038	7.064	0	2	23	regulation of cell proliferation
GO:0009653	0.04	1.654	13	19	925	anatomical structure morphogenesis
GO:0007076	0.04	36.806	0	1	3	mitotic chromosome condensation
GO:0007080	0.04	36.806	0	1	3	mitotic metaphase plate congression
GO:0014055	0.04	36.806	0	1	3	acetylcholine secretion
GO:0019722	0.04	36.806	0	1	3	calcium-mediated signaling
GO:0030953	0.04	36.806	0	1	3	spindle astral microtubule organization
GO:0032456	0.04	36.806	0	1	3	endocytic recycling
GO:0034508	0.04	36.806	0	1	3	centromere complex assembly
GO:0042694	0.04	36.806	0	1	3	muscle cell fate specification
GO:0045471	0.04	36.806	0	1	3	response to ethanol
GO:0045749	0.04	36.806	0	1	3	negative regulation of S phase of mitotic cell cycle
GO:0051310	0.04	36.806	0	1	3	metaphase plate congression
GO:0051382	0.04	36.806	0	1	3	kinetochore assembly
GO:0000226	0.043	3.19	1	4	98	microtubule cytoskeleton organization
GO:0006082	0.044	2.466	3	6	190	organic acid metabolic process
GO:0019752	0.044	2.466	3	6	190	carboxylic acid metabolic process
GO:0043436	0.044	2.466	3	6	190	oxoacid metabolic process
GO:0000003	0.046	1.48	25	33	1855	reproduction
GO:0003012	0.048	6.179	0	2	26	muscle system process
GO:0006936	0.048	6.179	0	2	26	muscle contraction
GO:0008283	0.048	6.179	0	2	26	cell proliferation
GO:0051728	0.048	6.179	0	2	26	cell cycle switching, mitotic to meiotic cell cycle
GO:0051729	0.048	6.179	0	2	26	germline cell cycle switching, mitotic to meiotic cell cycle
GO:0060184	0.048	6.179	0	2	26	cell cycle switching

APPENDIX I (continued)

111 Gene Subcluster Gene to GO Cellular Compartment test for over-representation

GOCCID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0005622	0	4.873	11	26	2148	intracellular
GO:0044428	0.001	10.843	0	4	82	nuclear part
GO:0044424	0.002	2.722	9	17	1659	intracellular part
GO:0042995	0.004	23.755	0	2	19	cell projection
GO:0000790	0.005	Inf	0	1	1	nuclear chromatin
GO:0001726	0.005	Inf	0	1	1	ruffle
GO:0005924	0.005	Inf	0	1	1	cell-substrate adherens junction
GO:0005925	0.005	Inf	0	1	1	focal adhesion
GO:0008091	0.005	Inf	0	1	1	spectrin
GO:0009288	0.005	Inf	0	1	1	bacterial-type flagellum
GO:0016323	0.005	Inf	0	1	1	basolateral plasma membrane
GO:0019861	0.005	Inf	0	1	1	flagellum
GO:0030055	0.005	Inf	0	1	1	cell-substrate junction
GO:0031252	0.005	Inf	0	1	1	cell leading edge
GO:0005954	0.01	197.154	0	1	2	calcium- and calmodulin-dependent protein kinase complex
GO:0030863	0.01	197.154	0	1	2	cortical cytoskeleton
GO:0030864	0.01	197.154	0	1	2	cortical actin cytoskeleton
GO:0044454	0.01	197.154	0	1	2	nuclear chromosome part
GO:0005634	0.011	2.64	5	10	872	nucleus
GO:0030054	0.012	13.003	0	2	33	cell junction
GO:0005654	0.014	12.212	0	2	35	nucleoplasm
GO:0044451	0.014	12.212	0	2	35	nucleoplasm part
GO:0031981	0.019	10.325	0	2	41	nuclear lumen
GO:0043229	0.02	2.202	7	13	1393	intracellular organelle
GO:0043226	0.02	2.2	7	13	1394	organelle
GO:0000228	0.021	65.701	0	1	4	nuclear chromosome
GO:0005912	0.021	65.701	0	1	4	adherens junction
GO:0008023	0.021	65.701	0	1	4	transcription elongation factor complex
GO:0070161	0.021	65.701	0	1	4	anchoring junction
GO:0043231	0.027	2.191	6	11	1146	intracellular membrane-bounded organelle
GO:0043227	0.028	2.188	6	11	1147	membrane-bounded organelle
GO:0043233	0.028	8.207	0	2	51	organelle lumen
GO:0070013	0.028	8.207	0	2	51	intracellular organelle lumen
GO:0043234	0.031	3.107	2	5	343	protein complex
GO:0031974	0.034	7.442	0	2	56	membrane-enclosed lumen
GO:0005669	0.036	32.838	0	1	7	transcription factor TFIID complex
GO:0044448	0.036	32.838	0	1	7	cell cortex part

APPENDIX I (continued)

GO:0005938	0.041	28.143	0	1	8	cell cortex
GO:0031461	0.041	28.143	0	1	8	cullin-RING ubiquitin ligase complex
GO:0044446	0.048	2.733	2	5	387	intracellular organelle part
GO:0044422	0.048	2.725	2	5	388	organelle part

111 Gene Subcluster Gene to GO Molecular Function test for over-representation

GOMFID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0005200	0	151.277	0	2	4	structural constituent of cytoskeleton
GO:0003723	0.003	5.829	1	5	141	RNA binding
GO:0003725	0.003	30.221	0	2	12	double-stranded RNA binding
GO:0004439	0.007	Inf	0	1	1	phosphatidylinositol-4,5-bisphosphate 5-phosphatase activity
GO:0018738	0.007	Inf	0	1	1	S-formylglutathione hydrolase activity
GO:0034593	0.007	Inf	0	1	1	phosphatidylinositol bisphosphate phosphatase activity
GO:0034595	0.007	Inf	0	1	1	phosphoinositide 5-phosphatase activity
GO:0004527	0.008	16.771	0	2	20	exonuclease activity
GO:0004649	0.014	148.146	0	1	2	poly(ADP-ribose) glycohydrolase activity
GO:0016944	0.014	148.146	0	1	2	RNA polymerase II transcription elongation factor activity
GO:0016788	0.016	3.247	2	6	299	hydrolase activity, acting on ester bonds
GO:0005515	0.019	1.925	14	21	2015	protein binding
GO:0003779	0.02	10.045	0	2	32	actin binding
GO:0003713	0.02	74.062	0	1	3	transcription coactivator activity
GO:0004652	0.02	74.062	0	1	3	polynucleotide adenylyltransferase activity
GO:0017056	0.02	74.062	0	1	3	structural constituent of nuclear pore
GO:0016813	0.027	49.368	0	1	4	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amidines
GO:0030145	0.034	37.021	0	1	5	manganese ion binding
GO:0070566	0.034	37.021	0	1	5	adenylyltransferase activity
GO:0016787	0.038	1.926	8	13	1136	hydrolase activity
GO:0003711	0.04	29.613	0	1	6	transcription elongation regulator activity
GO:0008092	0.041	6.683	0	2	47	cytoskeletal protein binding
GO:0042623	0.048	3.899	1	3	120	ATPase activity, coupled

111 Gene Subcluster Gene to GO Biological Process test for over-representation

GOBPID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0048522	0	9.924	1	7	106	positive regulation of cellular process
GO:0032502	0	3.096	27	42	3372	developmental process
GO:0048731	0	3.305	8	20	1017	system development

APPENDIX I (continued)

GO:0048856	0	2.987	11	24	1411	anatomical structure development
GO:0022403	0	5.714	2	9	235	cell cycle phase
GO:0010564	0	19.835	0	4	31	regulation of cell cycle process
GO:0007275	0	2.8	26	40	3251	multicellular organismal development
GO:0007126	0	6.459	1	7	158	meiosis
GO:0051327	0	6.416	1	7	159	M phase of meiotic cell cycle
GO:0048518	0	2.654	15	28	1916	positive regulation of biological process
GO:0051321	0	6.207	1	7	164	meiotic cell cycle
GO:0010605	0	9.414	1	5	77	negative regulation of macromolecule metabolic process
GO:0022402	0	4.628	2	9	286	cell cycle process
GO:0000279	0	5.118	2	8	228	M phase
GO:0009892	0	9.034	1	5	80	negative regulation of metabolic process
GO:0007049	0.001	4.289	2	9	307	cell cycle
GO:0031175	0.001	11.363	0	4	51	neuron projection development
GO:0051726	0.001	11.363	0	4	51	regulation of cell cycle
GO:0032501	0.001	2.472	29	41	3625	multicellular organismal process
GO:0051053	0.001	64.759	0	2	6	negative regulation of DNA metabolic process
GO:0048513	0.001	2.756	8	17	966	organ development
GO:0007548	0.001	2.999	6	14	710	sex differentiation
GO:0045927	0.001	2.395	14	25	1769	positive regulation of growth
GO:0035112	0.001	51.8	0	2	7	genitalia morphogenesis
GO:0048666	0.001	9.357	0	4	61	neuron development
GO:0040035	0.001	3.023	5	13	646	hermaphrodite genitalia development
GO:0048468	0.001	5.504	1	6	155	cell development
GO:0048806	0.001	2.997	5	13	651	genitalia development
GO:0003006	0.001	2.871	6	14	738	reproductive developmental process
GO:0009653	0.002	2.66	7	16	925	anatomical structure morphogenesis
GO:0040008	0.002	2.302	14	25	1822	regulation of growth
GO:0009887	0.002	3.571	3	9	364	organ morphogenesis
GO:0030154	0.002	4.29	2	7	232	cell differentiation
GO:0030182	0.002	7.831	1	4	72	neuron differentiation
GO:0051052	0.003	32.362	0	2	10	regulation of DNA metabolic process
GO:0045944	0.003	11.94	0	3	36	positive regulation of transcription from RNA polymerase II promoter
GO:0022414	0.003	2.454	8	17	1066	reproductive process
GO:0040007	0.003	2.144	19	29	2337	growth
GO:0045893	0.003	11.255	0	3	38	positive regulation of transcription, DNA-dependent
GO:0022008	0.003	7.19	1	4	78	neurogenesis

APPENDIX I (continued)

GO:0048699	0.003	7.19	1	4	78	generation of neurons
GO:0050793	0.003	3.945	2	7	251	regulation of developmental process
GO:0051254	0.004	10.941	0	3	39	positive regulation of RNA metabolic process
GO:0048869	0.004	3.912	2	7	253	cellular developmental process
GO:0048519	0.004	3.526	3	8	323	negative regulation of biological process
GO:0016043	0.004	2.888	4	11	555	cellular component organization
GO:0040010	0.004	2.183	13	22	1609	positive regulation of growth rate
GO:0040009	0.004	2.182	13	22	1610	regulation of growth rate
GO:0048729	0.004	3.435	3	8	331	tissue morphogenesis
GO:0045941	0.004	10.095	0	3	42	positive regulation of transcription
GO:0040020	0.005	23.526	0	2	13	regulation of meiosis
GO:0060341	0.005	23.526	0	2	13	regulation of cellular localization
GO:0010628	0.005	9.841	0	3	43	positive regulation of gene expression
GO:0030030	0.005	6.402	1	4	87	cell projection organization
GO:0009891	0.005	9.6	0	3	44	positive regulation of biosynthetic process
GO:0010557	0.005	9.6	0	3	44	positive regulation of macromolecule biosynthetic process
GO:0031328	0.005	9.6	0	3	44	positive regulation of cellular biosynthetic process
GO:0007399	0.005	6.325	1	4	88	nervous system development
GO:0045934	0.005	9.37	0	3	45	negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
GO:0051172	0.005	9.37	0	3	45	negative regulation of nitrogen compound metabolic process
GO:0007059	0.005	4.867	1	5	143	chromosome segregation
GO:0000003	0.005	2.092	15	24	1855	reproduction
GO:0051239	0.005	2.735	5	11	583	regulation of multicellular organismal process
GO:0045935	0.006	9.151	0	3	46	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
GO:0051173	0.006	9.151	0	3	46	positive regulation of nitrogen compound metabolic process
GO:0006366	0.006	6.105	1	4	91	transcription from RNA polymerase II promoter
GO:0040028	0.006	4.761	1	5	146	regulation of vulval development
GO:0048580	0.006	4.761	1	5	146	regulation of post-embryonic development
GO:0031325	0.006	8.942	0	3	47	positive regulation of cellular metabolic process

APPENDIX I (continued)

GO:0000904	0.006	8.742	0	3	48	cell morphogenesis involved in differentiation
GO:0048667	0.006	8.742	0	3	48	cell morphogenesis involved in neuron differentiation
GO:0048812	0.006	8.742	0	3	48	neuron projection morphogenesis
GO:0009893	0.007	8.368	0	3	50	positive regulation of metabolic process
GO:0010604	0.007	8.368	0	3	50	positive regulation of macromolecule metabolic process
GO:0045595	0.008	8.192	0	3	51	regulation of cell differentiation
GO:0048858	0.008	8.192	0	3	51	cell projection morphogenesis
GO:0009888	0.008	3.092	3	8	365	tissue development
GO:0051445	0.008	17.243	0	2	17	regulation of meiotic cell cycle
GO:0000076	0.008	Inf	0	1	1	DNA replication checkpoint
GO:0001934	0.008	Inf	0	1	1	positive regulation of protein amino acid phosphorylation
GO:0007063	0.008	Inf	0	1	1	regulation of sister chromatid cohesion
GO:0007064	0.008	Inf	0	1	1	mitotic sister chromatid cohesion
GO:0010562	0.008	Inf	0	1	1	positive regulation of phosphorus metabolic process
GO:0015807	0.008	Inf	0	1	1	L-amino acid transport
GO:0015812	0.008	Inf	0	1	1	gamma-aminobutyric acid transport
GO:0016185	0.008	Inf	0	1	1	synaptic vesicle budding from presynaptic membrane
GO:0016191	0.008	Inf	0	1	1	synaptic vesicle uncoating
GO:0016358	0.008	Inf	0	1	1	dendrite development
GO:0030174	0.008	Inf	0	1	1	regulation of DNA replication initiation
GO:0031573	0.008	Inf	0	1	1	intra-S DNA damage checkpoint
GO:0032297	0.008	Inf	0	1	1	negative regulation of DNA replication initiation
GO:0033044	0.008	Inf	0	1	1	regulation of chromosome organization
GO:0033564	0.008	Inf	0	1	1	anterior/posterior axon guidance
GO:0042327	0.008	Inf	0	1	1	positive regulation of phosphorylation
GO:0045005	0.008	Inf	0	1	1	maintenance of fidelity during DNA-dependent DNA replication
GO:0045128	0.008	Inf	0	1	1	negative regulation of reciprocal meiotic recombination
GO:0045835	0.008	Inf	0	1	1	negative regulation of meiosis
GO:0045937	0.008	Inf	0	1	1	positive regulation of phosphate metabolic process
GO:0048212	0.008	Inf	0	1	1	Golgi vesicle uncoating
GO:0048478	0.008	Inf	0	1	1	replication fork protection
GO:0048808	0.008	Inf	0	1	1	male genitalia morphogenesis

APPENDIX I (continued)

GO:0051177	0.008	Inf	0	1	1	meiotic sister chromatid cohesion
GO:0070142	0.008	Inf	0	1	1	synaptic vesicle budding
GO:0009790	0.008	1.952	20	29	2487	embryonic development
GO:0032990	0.009	7.707	0	3	54	cell part morphogenesis
GO:0031324	0.01	7.414	0	3	56	negative regulation of cellular metabolic process
GO:0002009	0.012	3.097	2	7	315	morphogenesis of an epithelium
GO:0060429	0.012	3.097	2	7	315	epithelium development
GO:0010629	0.014	6.435	1	3	64	negative regulation of gene expression
GO:0042127	0.014	12.306	0	2	23	regulation of cell proliferation
GO:0051049	0.014	12.306	0	2	23	regulation of transport
GO:0045132	0.015	4.52	1	4	121	meiotic chromosome segregation
GO:0045664	0.015	11.745	0	2	24	regulation of neuron differentiation
GO:0050767	0.015	11.745	0	2	24	regulation of neurogenesis
GO:0051960	0.015	11.745	0	2	24	regulation of nervous system development
GO:0060284	0.015	11.745	0	2	24	regulation of cell development
GO:0001539	0.016	127.298	0	1	2	ciliary or flagellar motility
GO:0006900	0.016	127.298	0	1	2	membrane budding
GO:0008156	0.016	127.298	0	1	2	negative regulation of DNA replication
GO:0031053	0.016	127.298	0	1	2	primary microRNA processing
GO:0031401	0.016	127.298	0	1	2	positive regulation of protein modification process
GO:0032270	0.016	127.298	0	1	2	positive regulation of cellular protein metabolic process
GO:0032387	0.016	127.298	0	1	2	negative regulation of intracellular transport
GO:0035196	0.016	127.298	0	1	2	gene silencing by miRNA, production of miRNAs
GO:0042001	0.016	127.298	0	1	2	hermaphrodite somatic sex determination
GO:0042308	0.016	127.298	0	1	2	negative regulation of protein import into nucleus
GO:0042992	0.016	127.298	0	1	2	negative regulation of transcription factor import into nucleus
GO:0046823	0.016	127.298	0	1	2	negative regulation of nucleocytoplasmic transport
GO:0048815	0.016	127.298	0	1	2	hermaphrodite genitalia morphogenesis
GO:0051224	0.016	127.298	0	1	2	negative regulation of protein transport
GO:0016441	0.016	11.233	0	2	25	posttranscriptional gene silencing
GO:0031047	0.016	11.233	0	2	25	gene silencing by RNA
GO:0035194	0.016	11.233	0	2	25	posttranscriptional gene silencing by RNA
GO:0008283	0.018	10.763	0	2	26	cell proliferation

APPENDIX I (continued)

GO:0051728	0.018	10.763	0	2	26	cell cycle switching, mitotic to meiotic cell cycle
GO:0051729	0.018	10.763	0	2	26	germline cell cycle switching, mitotic to meiotic cell cycle
GO:0060184	0.018	10.763	0	2	26	cell cycle switching
GO:0006996	0.018	2.827	3	7	343	organelle organization
GO:0040025	0.018	3.528	2	5	194	vulval development
GO:0006997	0.02	9.933	0	2	28	nucleus organization
GO:0000902	0.021	5.521	1	3	74	cell morphogenesis
GO:0048598	0.022	4.029	1	4	135	embryonic morphogenesis
GO:0006357	0.022	5.368	1	3	76	regulation of transcription from RNA polymerase II promoter
GO:0007067	0.023	9.221	0	2	30	mitosis
GO:0001956	0.024	63.64	0	1	3	positive regulation of neurotransmitter secretion
GO:0006368	0.024	63.64	0	1	3	RNA elongation from RNA polymerase II promoter
GO:0007032	0.024	63.64	0	1	3	endosome organization
GO:0007062	0.024	63.64	0	1	3	sister chromatid cohesion
GO:0010520	0.024	63.64	0	1	3	regulation of reciprocal meiotic recombination
GO:0016050	0.024	63.64	0	1	3	vesicle organization
GO:0031050	0.024	63.64	0	1	3	dsRNA fragmentation
GO:0031123	0.024	63.64	0	1	3	RNA 3'-end processing
GO:0032386	0.024	63.64	0	1	3	regulation of intracellular transport
GO:0033157	0.024	63.64	0	1	3	regulation of intracellular protein transport
GO:0035195	0.024	63.64	0	1	3	gene silencing by miRNA
GO:0042306	0.024	63.64	0	1	3	regulation of protein import into nucleus
GO:0042990	0.024	63.64	0	1	3	regulation of transcription factor import into nucleus
GO:0042991	0.024	63.64	0	1	3	transcription factor import into nucleus
GO:0043331	0.024	63.64	0	1	3	response to dsRNA
GO:0043631	0.024	63.64	0	1	3	RNA polyadenylation
GO:0045604	0.024	63.64	0	1	3	regulation of epidermal cell differentiation
GO:0045682	0.024	63.64	0	1	3	regulation of epidermis development
GO:0046822	0.024	63.64	0	1	3	regulation of nucleocytoplasmic transport
GO:0048488	0.024	63.64	0	1	3	synaptic vesicle endocytosis
GO:0051051	0.024	63.64	0	1	3	negative regulation of transport
GO:0051590	0.024	63.64	0	1	3	positive regulation of neurotransmitter transport

APPENDIX I (continued)

GO:0060631	0.024	63.64	0	1	3	regulation of meiosis I
GO:0009792	0.024	1.755	19	27	2434	embryonic development ending in birth or egg hatching
GO:0000280	0.025	8.901	0	2	31	nuclear division
GO:0016458	0.025	8.901	0	2	31	gene silencing
GO:0040018	0.026	2.843	2	6	289	positive regulation of multicellular organism growth
GO:0048285	0.026	8.604	0	2	32	organelle fission
GO:0051240	0.029	2.782	2	6	295	positive regulation of multicellular organismal process
GO:0040011	0.029	1.872	10	16	1243	locomotion
GO:0051094	0.03	4.715	1	3	86	positive regulation of developmental process
GO:0000087	0.031	7.818	0	2	35	M phase of mitotic cell cycle
GO:0022603	0.031	7.818	0	2	35	regulation of anatomical structure morphogenesis
GO:0001932	0.031	42.421	0	1	4	regulation of protein amino acid phosphorylation
GO:0006275	0.031	42.421	0	1	4	regulation of DNA replication
GO:0010720	0.031	42.421	0	1	4	positive regulation of cell development
GO:0031346	0.031	42.421	0	1	4	positive regulation of cell projection organization
GO:0043059	0.031	42.421	0	1	4	regulation of forward locomotion
GO:0045597	0.031	42.421	0	1	4	positive regulation of cell differentiation
GO:0045773	0.031	42.421	0	1	4	positive regulation of axon extension
GO:0045910	0.031	42.421	0	1	4	negative regulation of DNA recombination
GO:0045933	0.031	42.421	0	1	4	positive regulation of muscle contraction
GO:0050769	0.031	42.421	0	1	4	positive regulation of neurogenesis
GO:0050772	0.031	42.421	0	1	4	positive regulation of axonogenesis
GO:0051047	0.031	42.421	0	1	4	positive regulation of secretion
GO:0051983	0.031	42.421	0	1	4	regulation of chromosome segregation
GO:0051716	0.031	4.658	1	3	87	cellular response to stimulus
GO:0050789	0.033	1.683	26	33	3223	regulation of biological process
GO:0045137	0.033	4.548	1	3	89	development of primary sexual characteristics
GO:0032989	0.035	4.444	1	3	91	cellular component morphogenesis
GO:0010608	0.036	7.164	0	2	38	posttranscriptional regulation of gene expression
GO:0006302	0.039	31.811	0	1	5	double-strand break repair
GO:0006354	0.039	31.811	0	1	5	RNA elongation
GO:0007016	0.039	31.811	0	1	5	cytoskeletal anchoring at plasma membrane

APPENDIX I (continued)

GO:0009913	0.039	31.811	0	1	5	epidermal cell differentiation
GO:0018993	0.039	31.811	0	1	5	somatic sex determination
GO:0031399	0.039	31.811	0	1	5	regulation of protein modification process
GO:0033261	0.039	31.811	0	1	5	regulation of S phase
GO:0043056	0.039	31.811	0	1	5	forward locomotion
GO:0045814	0.039	31.811	0	1	5	negative regulation of gene expression, epigenetic
GO:0046716	0.039	31.811	0	1	5	muscle maintenance
GO:0051247	0.039	31.811	0	1	5	positive regulation of protein metabolic process
GO:0051320	0.039	31.811	0	1	5	S phase
GO:0009266	0.04	6.785	0	2	40	response to temperature stimulus
GO:0006793	0.04	2.118	5	9	588	phosphorus metabolic process
GO:0006796	0.04	2.118	5	9	588	phosphate metabolic process
GO:0065007	0.042	1.637	26	33	3273	biological regulation
GO:0000018	0.047	25.446	0	1	6	regulation of DNA recombination
GO:0000077	0.047	25.446	0	1	6	DNA damage checkpoint
GO:0030539	0.047	25.446	0	1	6	male genitalia development
GO:0031570	0.047	25.446	0	1	6	DNA integrity checkpoint
GO:0043058	0.047	25.446	0	1	6	regulation of backward locomotion
GO:0051050	0.047	25.446	0	1	6	positive regulation of transport
GO:0051223	0.047	25.446	0	1	6	regulation of protein transport
GO:0070201	0.047	25.446	0	1	6	regulation of establishment of protein localization
GO:0040029	0.047	6.135	0	2	44	regulation of gene expression, epigenetic
GO:0006897	0.049	5.992	0	2	45	endocytosis
GO:0010324	0.049	5.992	0	2	45	membrane invagination
GO:0010558	0.049	5.992	0	2	45	negative regulation of macromolecule biosynthetic process

19 Gene Subcluster Gene to GO Molecular Function test for over-representation

GOMFID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0004869	0.001	7158	0	1	2	cysteine-type endopeptidase inhibitor activity
GO:0004866	0.018	110.859	0	1	65	endopeptidase inhibitor activity
GO:0030414	0.019	105.851	0	1	68	peptidase inhibitor activity
GO:0004857	0.021	97.068	0	1	74	enzyme inhibitor activity
GO:0008237	0.037	52.827	0	1	134	metallopeptidase activity

363 Gene Subcluster Gene to GO Cellular Compartment test for over-representation

GOCCID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0016021	0.003	1.778	102	117	5496	integral to membrane

APPENDIX I (continued)

GO:0031224	0.004	1.772	102	117	5501	intrinsic to membrane
GO:0016020	0.005	1.762	107	120	5729	membrane
GO:0044425	0.008	1.682	104	117	5582	membrane part
GO:0005923	0.037	53.042	0	1	2	tight junction
GO:0016327	0.037	53.042	0	1	2	apicolateral plasma membrane
GO:0043296	0.037	53.042	0	1	2	apical junction complex
GO:0070160	0.037	53.042	0	1	2	occluding junction

363 Gene Subcluster Gene to GO Molecular Function test for over-representation

GOMFID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0004504	0	Inf	0	2	2	peptidylglycine monooxygenase activity
GO:0008519	0.003	32.321	0	2	6	ammonium transmembrane transporter activity
GO:0016705	0.004	7.086	1	4	41	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
GO:0015101	0.005	25.853	0	2	7	organic cation transmembrane transporter activity
GO:0016715	0.01	16.151	0	2	10	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced ascorbate as one donor, and incorporation of one atom of oxygen
GO:0003707	0.014	2.525	4	9	247	steroid hormone receptor activity
GO:0004768	0.016	Inf	0	1	1	stearoyl-CoA 9-desaturase activity
GO:0016756	0.016	Inf	0	1	1	glutathione gamma-glutamylcysteinyltransferase activity
GO:0016882	0.016	Inf	0	1	1	cyclo-ligase activity
GO:0030272	0.016	Inf	0	1	1	5-formyltetrahydrofolate cyclo-ligase activity
GO:0004879	0.016	2.461	4	9	253	ligand-dependent nuclear receptor activity
GO:0005507	0.022	9.932	0	2	15	copper ion binding
GO:0004857	0.027	3.728	1	4	74	enzyme inhibitor activity
GO:0000149	0.031	64.082	0	1	2	SNARE binding
GO:0008199	0.031	64.082	0	1	2	ferric iron binding
GO:0019905	0.031	64.082	0	1	2	syntaxin binding
GO:0003700	0.036	1.853	8	13	484	transcription factor activity
GO:0004365	0.046	32.036	0	1	3	glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) activity
GO:0005184	0.046	32.036	0	1	3	neuropeptide hormone activity

APPENDIX I (continued)

GO:0008943	0.046	32.036	0	1	3	glyceraldehyde-3-phosphate dehydrogenase activity

363 Gene Subcluster Gene to GO Biological Process test for over-representation

GOBPID	p-value	Log Odds Ratio	Expected Count	Count	Size	Term
GO:0000041	0.01	15.691	0	2	11	transition metal ion transport
GO:0006518	0.011	7.11	0	3	33	peptide metabolic process
GO:0032787	0.012	4.849	1	4	63	monocarboxylic acid metabolic process
GO:0009405	0.014	Inf	0	1	1	pathogenesis
GO:0017157	0.014	Inf	0	1	1	regulation of exocytosis
GO:0032876	0.014	Inf	0	1	1	negative regulation of DNA endoreduplication
GO:0042632	0.014	Inf	0	1	1	cholesterol homeostasis
GO:0046937	0.014	Inf	0	1	1	phytochelatin metabolic process
GO:0046938	0.014	Inf	0	1	1	phytochelatin biosynthetic process
GO:0055092	0.014	Inf	0	1	1	sterol homeostasis
GO:0019219	0.016	1.911	10	17	686	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
GO:0051171	0.016	1.908	10	17	687	regulation of nitrogen compound metabolic process
GO:0006887	0.026	8.817	0	2	18	exocytosis
GO:0031323	0.027	1.789	10	17	727	regulation of cellular metabolic process
GO:0001676	0.028	70	0	1	2	long-chain fatty acid metabolic process
GO:0008156	0.028	70	0	1	2	negative regulation of DNA replication
GO:0032875	0.028	70	0	1	2	regulation of DNA endoreduplication
GO:0042759	0.028	70	0	1	2	long-chain fatty acid biosynthetic process
GO:0045980	0.028	70	0	1	2	negative regulation of nucleotide metabolic process
GO:0055088	0.028	70	0	1	2	lipid homeostasis
GO:0006631	0.029	4.838	1	3	47	fatty acid metabolic process
GO:0007167	0.029	4.838	1	3	47	enzyme linked receptor protein signaling pathway
GO:0007179	0.032	7.836	0	2	20	transforming growth factor beta receptor signaling pathway
GO:0080090	0.033	1.743	11	17	744	regulation of primary metabolic process
GO:0010556	0.033	1.769	10	16	688	regulation of macromolecule biosynthetic process
GO:0009889	0.034	1.763	10	16	690	regulation of biosynthetic process
GO:0031326	0.034	1.763	10	16	690	regulation of cellular biosynthetic process
GO:0007178	0.035	7.422	0	2	21	transmembrane receptor protein serine/threonine kinase signaling pathway

APPENDIX I (continued)

GO:0048489	0.038	7.05	0	2	22	synaptic vesicle transport
GO:0006140	0.042	34.995	0	1	3	regulation of nucleotide metabolic process
GO:0006879	0.042	34.995	0	1	3	cellular iron ion homeostasis
GO:0009896	0.042	34.995	0	1	3	positive regulation of catabolic process
GO:0042023	0.042	34.995	0	1	3	DNA endoreduplication
GO:0042694	0.042	34.995	0	1	3	muscle cell fate specification
GO:0045732	0.042	34.995	0	1	3	positive regulation of protein catabolic process
GO:0055072	0.042	34.995	0	1	3	iron ion homeostasis
GO:0009408	0.049	6.128	0	2	25	response to heat

CITED LITERATURE

Velculescu, V., Zhang, L. Zhou, W., Vogelstein, J., Basrai, M., Bassett, D., Hieter, P., Vogelstein, B., and Kinzler, K. (1997). Characterization of the yeast transcriptome. Cell *88*, 243-51.

Affymetrix (2002). Datasheet, C. elegans and Drosophila Genome Arrays (pdf, 317 KB).

Alon, U. (2007). Network motifs: theory and experimental approaches. Nature Reviews. Genetics *8*, 450-61.

Baugh, L.R., Hill, A.A., Claggett, J.M., Hill-Harfe, K., Wen, J.C., Slonim, D.K., Brown, E.L., and Hunter, C.P. (2005). The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the C. elegans embryo. Development (Cambridge, England) *132*, 1843-54.

Baugh, L.R., Hill, A.A., Slonim, D.K., Brown, E.L., and Hunter, C.P. (2003). Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. Development (Cambridge, England) *130*, 889-900.

Benjamini, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal Of The Royal Statistical Society Series B.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics (Oxford, England) *19*, 185-93.

Borok, M.J., Tran, D.A., Ho, M.C.W., and Drewell, R.A. (2009). Dissecting the regulatory switches of development: lessons from enhancer evolution in Drosophila. Development *137*, 5-13.

Brenner, S. (1974). The genetics of Caenorhabditis elegans. Genetics *77*, 71-94.

Brown, C.T. (2008). Computational Approaches to Finding and Analyzing cis-Regulatory Elements. p. 1-29.

Cameron, R.A., and Davidson, E.H. (2009). Flexibility of transcription factor target site position in conserved cis-regulatory modules. Developmental Biology *336*, 122-135.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. Blood *103*, 2771-8.

Coll, M., Seidman, J.G., and Müller, C.W. (2002). Structure of the DNA-bound T-box domain of human TBX3, a transcription factor responsible for ulnar-mammary syndrome. Structure (London, England : 1993) *10*, 343-56.

Crooks, G.E., Hon, G., Chandonia, J., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Research *14*, 1188-90.

Ganter, B., and Lipsick, J.S. (1999). Myb and oncogenesis. Advances In Cancer Research *76*, 21-60.

Gaudet, J., and Mango, S.E. (2002). Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4. Science (New York, N.Y.) *295*, 821-5.

Gentleman, R.C., Carey, V.J., Bates, D.M., Ben Bolstad, Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biology *5*, 1-16.

Gerstein, M.B., Lu, Z.J., van den Van Nostrand, E.L., van der Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., et al. (2010). Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. Science *330*, 1775-1787.

Ghosh, T.K., Packham, E.A., Bonser, A.J., Robinson, T.E., Cross, S.J., and Brook, J.D. (2001). Characterization of the TBX5 binding site and analysis of mutations that cause Holt-Oram syndrome. Human Molecular Genetics *10*, 1983-94.

Habets, P.E., Moorman, A.F., Clout, D.E., van Roon, M.A., Lingbeek, M., van Lohuizen, M., Campione, M., and Christoffels, V.M. (2002). Cooperative action of Tbx2 and Nkx2.5 inhibits ANF expression in the atrioventricular canal: implications for cardiac chamber formation. Genes &Amp;Amp; Development *16*, 1234-46.

Harris, T.W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W.J., La Cruz, de, N., Davis, P., Duesbury, M., Fang, R., et al. (2010). WormBase: a comprehensive resource for nematode research. Nucleic Acids Research *38*, D463-7.

Hirose, T., Galvin, B.D., and Horvitz, H.R. (2010). Six and Eya promote apoptosis through direct transcriptional activation of the proapoptotic BH3-only gene egl-1 in Caenorhabditis elegans. Proceedings Of The National Academy Of Sciences Of The United States Of America *107*, 15479-84.

Hoogaars, W.M., Barnett, P., Moorman, A.F., and Christoffels, V.M. (2007). T-box factors determine cardiac design. Cellular And Molecular Life Sciences : CMLS *64*, 646-60.

Kal, A.J., van Zonneveld, A.J., Benes, V., van den Berg, M., Koerkamp, M.G., Albermann, K., Strack, N., Ruijter, J.M., Richter, A., Dujon, B., et al. (1999). Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. Molecular Biology Of The Cell *10*, 1859-72.

Kato, M., and Tsunoda, T. (2001). Lag analysis of genetic networks in the cell cycle of budding yeast. GENOME INFORMATICS SERIES.

Kispert, A., and Hermann, B.G. (1993). The Brachyury gene encodes a novel DNA binding protein. The EMBO Journal *12*, 4898-9.

Linask, K.K., Han, M.D., Artman, M., and Ludwig, C.A. (2001). Sodium-calcium exchanger (NCX-1) and calcium modulation: NCX protein expression patterns and regulation of early heart development. Developmental Dynamics : An Official Publication Of The American Association Of Anatomists *221*, 249-64.

Lingbeek, M.E., Jacobs, J.J., and van Lohuizen, M. (2002). The T-box repressors TBX2 and TBX3 specifically regulate the tumor suppressor gene p14ARF via a variant T-site in the initiator. The Journal Of Biological Chemistry *277*, 26120-7.

Liu, H., Brannon, A.R., Reddy, A.R., Alexe, G., Seiler, M.W., Arreola, A., Oza, J.H., Yao, M., Juan, D., Liou, L.S., et al. (2010). Identifying mRNA targets of microRNA dysregulated in cancer: with application to clear cell Renal Cell Carcinoma. BMC Systems Biology *4*, 51-17.

Loots, G.G., and Ovcharenko, I. (2007). Mulan: multiple-sequence alignment to predict functional elements in genomic sequences. Methods In Molecular Biology (Clifton, N.J.) *395*, 237-54.

Macindoe, I., Glockner, L., Vukašin, P., Stennard, F.A., Costa, M.W., Harvey, R.P., Mackay, J.P., and Sunde, M. (2009). Conformational Stability and DNA Binding Specificity of the Cardiac T-Box Transcription Factor Tbx20. Journal Of Molecular Biology *389*, 606-618.

Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. Nature Reviews. Genetics *12*, 671-82.

McKay, S.J., Johnsen, R., Khattra, J., Asano, J., Baillie, D.L., Chan, S., Dube, N., Fang, L., Goszczynski, B., Ha, E., et al. (2003). Gene expression profiling of cells, tissues, and developmental stages of the nematode C. elegans. Cold Spring Harbor Symposia On Quantitative Biology *68*, 159-69.

Mello, C.C., Kramer, J.M., Stinchcomb, D., and Ambros, V. (1991). Efficient gene transfer in C.elegans: extrachromosomal maintenance and integration of transforming sequences. The EMBO Journal *10*, 3959-70.

Metzstein, M.M., Hengartner, M.O., Tsung, N., Ellis, R.E., and Horvitz, H.R. (1996). Transcriptional regulator of programmed cell death encoded by Caenorhabditis elegans gene ces-2. Nature *382*, 545-7.

Minguillon, C., and Logan, M. (2003). The comparative genomics of T-box genes. Briefings In Functional Genomics And Proteomics *2*, 224-33.

Murie, C., Woody, O., Lee, A.Y., and Nadon, R. (2009). Comparison of small n statistical tests of differential expression applied to microarrays. BMC Bioinformatics *10*, 45.

Müller, C.W., and Herrmann, B.G. (1997). Crystallographic structure of the T domain-DNA complex of the Brachyury transcription factor. Nature *389*, 884-8.

Niu, W., Lu, Z.J., Zhong, M., Sarov, M., Murray, J.I., Brdlik, C.M., Janette, J., Chen, C., Alves, P., Preston, E., et al. (2011). Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans. Genome Research *21*, 245-254.

Pflugfelder, G. (2009). omb and circumstance. Journal Of Neurogenetics *23*, 15-33.

Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W., and Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Research *38*, D105-10.

Qian, L., Mohapatra, B., Akasaka, T., Liu, J., Ocorr, K., Towbin, J.A., and Bodmer, R. (2008). Transcription factor neuromancer/TBX20 is required for cardiac function in Drosophila with implications for human heart disease. Proceedings Of The National Academy Of Sciences Of The United States Of America *105*, 19833-8.

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

Roy Chowdhuri, S., Crum, T., Woollard, A., Aslam, S., and Okkema, P.G. (2006). The T-box factor TBX-2 and the SUMO conjugating enzyme UBC-9 are required for ABa-derived pharyngeal muscle in C. elegans. Developmental Biology *295*, 664-677.

Ryu, J., Najand, N., and Brook, W.J. (2011). Tinman is a direct activator of midline in the Drosophila dorsal vessel. Developmental Dynamics : An Official Publication Of The American Association Of Anatomists *240*, 86-95.

Showell, C., Binder, O., and Conlon, F.L. (2004). T-box genes in early embryogenesis. Developmental Dynamics : An Official Publication Of The American Association Of Anatomists *229*, 201-18.

Smith, P.A., and Mango, S.E. (2007). Role of T-box gene tbx-2 for anterior foregut muscle development in C. elegans. Developmental Biology *302*, 25-39.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical Applications In Genetics And Molecular Biology *3*, Article3-28.

Stennard, F., and Harvey, R. (2005). T-box transcription factors and their roles in regulatory hierarchies in the developing heart. Development (Cambridge, England) *132*, 1-14.

Stirnimann, C.U., Ptchelkine, D., Grimm, C., and Müller, C.W. (2010). Structural basis of TBX5-DNA recognition: the T-box domain in its DNA-bound and -unbound form. Journal Of Molecular Biology *400*, 71-81.

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

Thompson, J.D., Gibson, T.J., and Higgins, des, G. (2002). Multiple sequence alignment using ClustalW and ClustalX. Current Protocols In Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [Et Al.] *Chapter 2*, Unit 2.3.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. Science (New York, N.Y.) *270*, 484-7.

Warrens, A.N., Jones, M.D., and Lechler, R.I. (1997). Splicing by overlap extension by PCR using asymmetric amplification: an improved technique for the generation of hybrid proteins of immunological interest. Gene *186*, 29-35.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nature Reviews. Genetics *5*, 276-87.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics (Oxford, England) *25*, 1189-91.

Zaragoza, von, M., Lewis, L.E., Sun, G., Wang, E., Li, L., Said-Salman, I., Feucht, L., and Huang, T. (2004). Identification of the TBX5 transactivating domain and the nuclear localization signal. Gene *330*, 9-18.

Zhong, M., Niu, W., Lu, Z.J., Sarov, M., Murray, J.I., Janette, J., Raha, D., Sheaffer, K.L., Lam, H.Y.K., Preston, E., et al. (2010). Genome-Wide Identification of Binding Sites Defines Distinct Functions for Caenorhabditis elegans PHA-4/FOXA in Development and Environmental Response. Plos Genetics *6*, e1000848.

VITA

NAME Thomas J. Ronan III

EDUCATION

University of Chicago, B.A. History, 1995
Loyola University School of Law, J.D., 2000
University of Missouri, B.A. Chemistry, 2003

PUBLICATIONS

1. Piriyapongsa, J., Jordan, I. K., Conley, A. B., Ronan, T., and Smalheiser, N. R. (2011) Transcription factor binding sites are highly enriched within microRNA precursor sequences. Biology Direct, 6:61.

PUBLISHED ABSTRACTS

1. Ronan, T., Clary, L., and Okkema, P. G. Leveraging existing genomic and microarray data to find direct targets of the *C. elegans* transcription factor TBX-2. (2011) 18th International *C. elegans* Meeting. University of California, Los Angeles, CA
2. Clary, L., Ronan, T., and Okkema, P. G. *C. elegans* TBX-2 is a SUMOylation dependent transcriptional repressor. (2011) 18th International *C. elegans* Meeting. University of California, Los Angeles, CA
3. Williams, A., Orozco-Nunnelly, D., Mezzich, R., Mohammad, D., Ronan, T. and Warpeha, K. Transformation of Human Pirin into Arabadopsis Thaliana Plants and Effects on Seedling Development. (2012) University of Illinois at Chicago Student Research Forum. University of Illinois at Chicago, Chicago, IL

HONORS, AWARDS, AND SERVICE

Funded Research Assistantship, Warpeha Lab	2011 – 2012
Funded Research Assistantship, Okkema Lab	2009 – 2011
Board of Trustees Tuition Waiver	2008 – 2009
Reviewer, University of Illinois Student Bioengineering Journal	2010

TEACHING EXPERIENCE

Law School Admissions Test (LSAT) Master Tutor	2002-2005
Medical College Admissions Test (MCAT) Chemistry and Verbal Reasoning Instructor	2003-2005

PROFESSIONAL MEMBERSHIPS

Genetics Society of America	2011
Admitted to the practice of law before the Missouri Bar	2004